AN APPLICATION OF GENERALIZABILITY THEORY TO THE ASSESSMENT OF WRITING ABILITY

By

MARIA MAGDALENA LLABRE

A DISSERTATION PRESENTED TO THE GRADUATE COUNCIL OF
THE UNIVERSITY OF FLORIDA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1978

ACKNOWLEDGEMENTS

The members of my Doctoral Committee deserve special recognition for their assistance with this dissertation. The chairman of my committee, Dr. William B. Ware, has my deepest respect and admiration. His standard of excellence has served as a model for me. To him I am indebted for providing innumerable opportunities for learning. Dr. Linda M. Crocker has also been most influential during my graduate program. I appreciate her sound advice and consistent encouragement. My sincere gratitude goes to Dr. Ramon C. Littell for the support he has given me along with many explanations of statistical methods. I also appreciate the continuous guidance of Dr. John M. Newell.

I would also like to thank Dr. James H. Goodnight of the SAS Institute for his invaluable assistance with the data analysis.

To my friends Mary Lynn, Barbara Boss, and Shirley Bowes, I am grateful for all the hours they spent rating the compositions without losing their sense of humor. The help of Richard Thompson in facilitating the analysis is gratefully recognized. Special thanks go to Louise Stephenson for typing this manuscript.

Finally, I acknowledge the personal support of my husband, Brainard, whose encouragement and understanding have given me the strength to carry on.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

Abstract of Dissertation Presented to the Graduate Council
of the University of Florida in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

AN APPLICATION OF GENERALIZABILITY THEORY TO THE ASSESSMENT OF WRITING ABILITY

By

MARIA MAGDALENA LLABRE

August 1978

Chairman: William B. Ware
Major Department: Foundations of Education

Classical reliability theory, as used in the social sciences, has been restricted by a model which specifies one undifferentiated error component. This restriction has limited the applicability of the model and has obscured its interpretation. Recent advancements in psychometric theory provide more flexible models which permit the investigation of multiple sources of error variation. Under the rubric of generalizability theory, these methods are based on R. A. Fisher's work on the analysis of variance and the factorial experiment.

Generalizability theory is potentially very useful in many areas of research suffering from inconsistency of measurement. In particular, the theory is applicable to the assessment of writing ability from written compositions. However, applied studies in this area are lacking.

The literature on the measurement of writing ability has identified several sources of error affecting the reliability of written compositions.

The most common sources of error noted are inconsistency across raters, modes, and occasions. In spite of the recognition of these sources of variation, most researchers who have studied the reliability of written composition have examined the issue only in terms of inter-rater reliability. Implicit in the concept of inter-rater reliability is the assumption that fluctuations among raters are the only errors in the model. This study incorporated three facets: raters, modes, and occasions, in a split-plot factorial design in order to examine the results obtained by taking into account more than one source of error through the methodology of generalizability theory.

Samples of writing from 104 fourth graders were obtained under selected mode and occasion conditions. Each sample was scored by four trained raters. In the design, the students were considered as nested within a higher classification, the classes. The number of students in each class was not constant. Therefore, this study also extended the principles of generalizability theory to unbalanced designs.

Point estimates of the variance components for all effects in the model were obtained through the MIVQUE method. Negative estimates were replaced by zeros. The relative magnitude of the estimates indicated that students could be differentiated on the basis of their ratings. However, the classes as units could not be distinguished. The estimates also showed that errors resulting from variability in the quality of writing across occasions and modes outweigh those stemming from differences among raters. Furthermore, occasions represented a greater source of error than modes. With training and practice, raters can consistently score the writing samples of students using a general impression method.

Assuming homogeneity of variance, unbiased generalizability coefficients were obtained for seven universes of generalization. These universes represented generalization across one facet, two facets, or all three facets simultaneously. The coefficients indicated that, to obtain acceptable levels of generalizability, at least six samples of writing from each person are necessary.

The standard error of measurement which may be used in constructing confidence intervals around a person's universe score was also examined. The results from this examination paralleled those based on the generalizability coefficients.

A supplementary analysis which allowed a comparison of the estimates obtained through the MIVQUE method to those derived using expected mean squares, resulted in similar values for all estimates in a model without the classes effect. These results were interpreted as lending support to the MIVQUE method.

It was concluded that generalizability theory is very useful for clarifying problems in estimating reliability in the area of writing ability. Furthermore, the theory need not be limited to situations with balanced data. Valid methods of variance component estimation documented in the statistical literature may be used with unbalanced designs.

CHAPTER I

INTRODUCTION

The concept of reliability in educational research has undergone notable refinements with a resulting increase in clarity and applicability. However, these conceptual developments have not been matched by applications in many content areas. For example, the reliability of essay tests still represents a confusing issue, partly because of the continued use of the classical model for its investigation. This study represents an attempt to "bridge the gap" between some recognized methodological needs in the field of written language arts and advancements in measurement theory.

Classical reliability theory has been based on a model (originated by Spearman in 1904) which states that a person's observed score is the sum of a true score component and an undifferentiated error component as shown below:

$$(1) \quad X = \pi + e \ .$$

The true and error components are assumed to be independent of each other. Therefore, the variance of the observed scores for a group of individuals can be partitioned into the sum of independent variance components as shown in equation (2).

$$(2) \quad \sigma_x^2 = \sigma_\pi^2 + \sigma_e^2 \ .$$

The reliability of a test is then defined as the ratio of the true score variance to observed score variance.

$$(3) \quad r_{xx} = \frac{\sigma_\pi^2}{\sigma_x^2} \ .$$

Since $\sigma_{\P}^2$ and σ_e^2 are unknown, in practice reliability is estimated by computing the correlation between parallel forms of the test. In order for tests to be parallel, they must have equal means, equal variances, and equal intercorrelations among items. From these restrictions and the assumptions imposed on the model (1), it can be shown that if two tests are parallel, their correlation equals (3) above (for proof see Magnusson, 1967).

In addition to the restriction of parallelism, classical theory considers the error component to be undifferentiated, that is, various sources of inconsistency which may affect the reliability of the test are grouped together in a single error term. Different procedures for constructing parallel tests (e.g. test-retest, split-half) make different assumptions about what constitutes the source of error in the model. Therefore, following the classical model, more than one interpretation of the same error component is possible.

The limitations of the classical model mentioned above render it inefficient in many real life situations for several reasons. First, the condition of parallelism is seldom met in the real world. It is common to find that supposedly parallel tests have different means. When tests have different means, the formulas which assume equality provide an underestimate of the reliability (Ebel, 1951). Second, by including only one error component which changes in meaning depending on the method of obtaining parallel forms of the test, the classical model can lead to some confusion. Unless the type of coefficient is reported, the model provides no clues for the interpretation of the error component. Finally, when more than one coefficient is desired

under the classical model, more than one study must be conducted. As a result, the model does not allow for the consideration of error resulting from interactions among sources.

In order to overcome these deficits inherent in the classical model, some measurement specialists have adopted R. A. Fisher's conceptualization of the factorial experiment, a method of classifying observations along more than one dimension; and the analysis of variance, a procedure which partitions total variability into identifiable sources. These two powerful tools have allowed for the possibility of releasing the restriction of parallelism and have provided a systematic approach to the simultaneous consideration of multiple sources of error variation.

The applicability of these concepts to social science research and specifically to the reliability problem was explicitly discussed by Lindquist (1953). Since then, these techniques have been widely used in testing hypotheses about group differences but only rarely in assessing reliability.

More recently, Cronbach and his colleagues have assembled all of the work which has been done along these lines under the rubric of generalizability theory. The synthesis of their efforts is described in their 1972 book entitled The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Basically, generalizability theory uses the analysis of variance approach in the estimation of reliability. Rather than emphasizing the computation of reliability coefficients as the classical theory does, the emphasis is on the estimation of variance components for all identifiable sources incorporated into the design. The theory allows for unequal means, decomposes the error term into separate sources, and requires the explicit consideration of the factors identifying the population of measures being studied.

If desired, the variance components can be used in the computation of "generalizability coefficients." These are intraclass correlations analogous to reliability coefficients. Within a less restrictive model which partitions the variance into several sources, more than one coefficient is possible from just one study. One of the most important advantages of this approach is that the analysis of variance technique can be applied to many different types of experimental designs. When the levels of the factors included in the design result from a factorial experiment, the analysis of variance can provide estimates of the variability due to interactions among factors. As was previously noted, these interactions were undetectable under the classical model.

The Terminology of Generalizability Theory

Generalizability theory is considered by its developers as an extension and liberalization of classical reliability theory. An important distinction is made between two kinds of studies: G and D. A G-study or generalizability study is one where the sources and magnitude of the variability in one particular measurement instrument are investigated. A G-study is analogous to a reliability study in the classical sense.

A D-study or decision study is one which uses information concerning the generalizability of a specific measurement tool for decision making purposes. Two types of decisions are identified: absolute or comparative. Absolute decisions are those which consider each individual separately. Placement and classification decisions are both absolute decisions. A specific example would be a decision made by a guidance counselor to place a student in one of several curriculum programs on the basis of the student's score on a test. A comparative decision is based on a

comparison of one individual to another or a comparison among groups of individuals. Selection decisions, as well as decisions involving group differences, fall under this category. An example of a comparative decision occurs when the scores on a test are used as the dependent variable for comparing the performance of two groups participating in an experiment.

In generalizability theory an observation is considered to be a sample from the total universe of observations which could have been made. The observation is described in terms of the conditions under which it is made. Two or more conditions of the same type constitute a facet. With one exception, in Fisherian terms a facet is a factor; conditions are simply the levels of a factor. The exception is that "persons" is never considered a facet in a G-study even though it is a factor.

When conducting a G-study, the investigator should include as many of the facets which are considered to affect the reliability of the measure as possible. From each facet, the investigator samples a set of conditions under a particular design. The observations are then made under the set of conditions sampled. The set of all possible observations which could be included in the G-study is referred to as the universe of admissible observations.

When the sampling of conditions is done at random and the universe of conditions is sufficiently large, the investigator is operating under a random effects model. This model is the one most commonly used in the context of generalizability theory although fixed and mixed models are also possible.

Regardless of the model used, the point estimates of the variance components are obtained by computing the mean squares from the analysis of variance and setting them equal to their corresponding expected mean squares. A helpful but unnecessary restriction has been made in generalizability theory that equal numbers of observations appear in the subclassifications of the design. This restriction simplifies the procedure for obtaining point estimates of the variance components but is not absolutely necessary. With equal numbers of observations, the mean squares from the analysis of variance are unique and are "the best" estimates possible. Having unequal numbers of observations creates a situation where the investigator must decide which of several sums of squares (and, therefore, mean squares) to use. In either case, the expected values of the sums of squares are linear combinations of the variance components. Therefore, solving for a set of simultaneous equations will result in point estimates.

The estimates of the variance components obtained under a G-study can be used in subsequent D-studies as long as the facets included in the D-study were also included in the G-study. The conditions, however, do not have to be the same if a random effects model is being considered.

The set of all possible observations to which an investigator carrying out a D-study wishes to generalize is termed the universe of generalization. This universe must then be a subset of the universe of admissible observations of the G-study providing the variance component estimates. This relationship between a G-study and subsequent D-studies implies that the utility of a G-study depends upon its ability to provide estimates of as many components of variance as might arise in future D-studies. That is, a G-study providing estimates of three components of

variance is more useful than one where only one of those three components is estimable, all other things being equal.

Purpose of the Study

The purpose of this study was to apply the principles of generalizability theory to the assessment of written composition. The significance of this study is twofold. First, it illustrates how theoretical measurement concepts can be applied and extended to fit specific problems encountered in assessment. Second, it provides guidelines to applied researchers and evaluators in the field of writing for improved methods of estimating the reliability of their assessment procedures.

With the current movement toward teaching basic skills, the effectiveness of tests in assessing progress in reading, writing, and arithmetic is under scrutiny. Of these three areas, writing presents a paradoxical conflict. While objective tests of writing ability are generally more reliable, essay tests or written compositions are considered to be more valid measures of writing ability (Coffman, 1971). The opinion of most specialists in the field of language arts is that the validity of essay tests should not be traded for the higher reliability of objective tests (McColly, 1970).

Given this preference for the essay test in the assessment of writing skill, any efforts to improve the quality of measurement in this area should focus on this test form. Unfortunately, advancements in measurement theory and practice have been, for the most part, restricted to objective tests. However, generalizability theory offers great potential usefulness for upgrading the reliability of measures of written composition. Applied studies are needed to test the reality of that potential.

Statement of the Problem

The need to study the application of generalizability theory to the assessment of writing skills becomes apparent when the recommendations on research methodology from leading curriculum specialists in written language arts are examined. Although the specific recommendations will be discussed in the following chapter, at this point we will note that more than one source of variability affecting the reliability of written composition has been identified. The most common sources of error noted are inconsistency across raters, modes, and occasions.

In spite of the recognition of these sources of variation, most researchers who have studied the reliability of written composition have examined the issue only in terms of inter-rater reliability. Implicit to the concept of inter-rater reliability is the assumption that fluctuations among raters is the only source of error in the model. This study incorporated three facets in a split-plot factorial design in order to examine the results obtained by taking into account more than one source of error. Following the recommendations of Brennan (1975), the students were seen as nested in a higher classification, the classes. The number of students in each class was not constant. Therefore, this study also explored procedures for obtaining estimates of the variance components which are applicable to unbalanced designs. An unbalanced design as defined here is one with unequal numbers of observations in the subclassifications (Searle, 1971a).

Using the results from this study, we will be able to assess the magnitude of each source of variability and determine which ones are the most important to control in order to obtain reliable assessments

of writing skill. Based on these results, we will be able to make recommendations for the design of future D-studies using the same method of assessment. These recommendations will include the nature of the facets which must be considered as well as the frequency with which each facet should be sampled. Both absolute and comparative decisions will be taken into account.

In addition, the estimates of the variance components will be used in the computation of several generalizability coefficients. The coefficients to be considered are those which provide estimates of the reliability when generalization is intended in either one dimension (across raters, modes, or occasions), two dimensions (raters and modes, etc.), or three dimensions (raters, modes, and occasions).

Significance of the Study

A renewed nationwide interest in the assessment of writing may be evidenced by the following events:

1. A compositional writing subtest is being reinstated on the Scholastic Aptitude Test (SAT) examination used by many colleges and universities for student selection and placement.

2. Interest in expanding the base of knowledge on the writing process has been underscored by National Institute of Education (NIE) in the 1978 competition for Basic Skills Awards.

3. A number of states now include writing as a skill to be tested in their efforts to establish statewide standards for minimum educational competency.

Those responsible for the preparation of these examinations will naturally be governed by practical considerations, such as the demonstrable quality of those examinations. Demonstrating the reliability

of their techniques must be one of the considerations. Generalizability coefficients, which take into account various sources of error associated with writing assessment, provide unambiguous estimates of reliability. As a result, they are preferable to the traditional inter-rater correlation coefficient.

CHAPTER II

REVIEW OF THE LITERATURE

The literature reviewed in this chapter has been selected from three distinct fields: language arts, measurement theory, and statistical methodology. The review is organized in the following manner. First, selected literature pertinent to the assessment of writing ability is presented to establish the rationale for the content area of this study. Particular attention will be given to studies involving primary grade children. Next, the development of generalizability theory is traced, followed by references illustrating applications of the theory. Finally, selected references from the literature on methods of variance component estimation are reviewed with emphasis on methods that are applicable to unbalanced designs. These designs have greatest utility in determining generalizability coefficients for assessments of written composition.

The Assessment of Writing Ability

Teachers and nonteachers alike would agree that writing is one of the most important subjects taught in schools. But the importance of the subject has not been accompanied by effective assessment. Evaluating students' writing performance continues to be a problem for writing specialists, English teachers, and researchers investigating this complex area. Both objective tests and compositional writing (essay tests) continue to be used (Coffman, 1971). However, the balance seems to be on the side of essay tests. After reviewing several standardized

objective tests of writing, McCaig (1977) recommended "to evaluate achievement in writing, evaluate the writing of children"(p.491).

Several other experts in the field also agree that writing ability is best determined by looking at actual writing performance (Coffman, 1971 & McColly, 1970). The members of a recent Louisiana State Department of Education conference on minimum writing proficiency unanimously recommended that any test of writing proficiency include a sample of the student's writing (Suhor, 1977).

Objective tests are generally not recommended. A quote from Braddock (1976) emphasizes the point:

> At this stage of our understanding of writing and of testing, it is difficult to believe that any standardized test will be constructed which can measure such ability. Therefore, anyone who professes to evaluate "writing ability" with a standardized test is either telling a falsehood or speaking from ignorance. (p.119)

At the present time, essay tests are included in a number of commonly used tests of English. Examples of these are the Language Skills Examination, the College Entrance Examination Board, and the writing test developed by NAEP. These may be used for the prediction of success in English, placement in special courses, exemption from required courses, program evaluation, and experimental or correlational research (Cooper and Odell, 1977).

Sources of Error in Essay Tests

The problem of the reliability of essay tests has been widely recognized for some time (Meckel, 1963). Adequate reliability is particularly important in required writing courses in which students must earn a satisfactory grade and also in research, when essay tests

are used as a measure of gains or losses in skill which are to be attributed to experiments in teaching methods.

Diederich (1957) suggested that the major problem of grading essays has to do with variation in the grades assigned by different readers. Commenting on the difficulties involved in grading such tests, he pointed out that when 10 readers read a set of papers without discussing standards, it is likely that average papers will receive the whole range of grades. He suggested three criteria for judging essay tests of writing ability. First, the writing assignment should be like the writing students do in the normal course of events. Second, the grading should be independent of the writer's knowledge of the subject matter. Finally, the topic must be within the student's comprehension. These criteria were met in the selection of assignment and in the grading of the samples used in this study.

To improve the reliability of essays, he recommended that all students write on the same topic, that readers be trained, and that at least two samples of writing be obtained from each student. This last recommendation suggests a second source of variation related to the reliability problem. Meckel was aware of this source when he said: "samples of writing done over a semester are obviously a better index of writing ability than a single essay" (p.988).

Braddock, Lloyd-Jones, and Schoer (1963), after screening and reviewing 484 studies on writing, discussed four sources of variation which should be taken into account when rating compositions. These sources are: the writer variable, the assignment variable, the rater variable, and the colleague variable. The writer variable refers to day-to-day fluctuations in the writing performance of individuals,

particularly the performance of better writers. On this issue these authors recommend that each student write at least twice.

Under the assignment variable, Braddock et al. included four aspects: topic, mode, time, and situation. They hypothesized that variation in mode may have a stronger effect on the quality of writing than variation in topic. The modes considered by these authors were: narration, description, exposition, argument, and criticism. With respect to time and condition, their recommendation was to allow as much as 20 to 30 minutes of writing time for primary grade children and to standardize the conditions across all children.

The rater variable, as defined by Braddock et al., refers to the tendency of a rater to vary in his/her own standards of evaluation while the colleague variable refers to variation in standards across different raters. The existence of inter-rater variability has been substantiated very frequently by research. Braddock et al. recommended that the raters have a common set of criteria and that they practice together in applying those criteria consistently. Two additional recommendations were offered in order to reduce the inter-rater variation. One of them was to preserve the anonymity of the writer. (These recommendations were previously made by Diederich). The second one was to control for rater fatigue. As will be shown in the next chapter, these recommendations were followed in the rating of the samples used in this study.

McColly (1970) categorized the sources of error in grading essay tests of writing ability into three general sources: students, readers, and topics. In determining his classification scheme, he considered the categories offered by Braddock et al. as well as those proposed by French (1962). French's categories, almost identical to McColly's,

consist of student errors, test errors (the task and the topic), and scale errors (reader disagreement).

Under the student source, McColly considered conditions such as distractions (both internal and external) as well as the motivation of the student. He recommended allowing the student at least 40 to 45 minutes of writing time.

With respect to readers, McColly concurred that readers must be given the proper training and orientation as well as the opportunity to practice. Practice is indispensable in establishing the proper speed and rate. He makes the following general statement in this regard: "up to the point where the prose becomes ununderstandable, the faster the rate and speed, the more valid and reliable the judgement"(p.150).

As far as the topic is concerned, McColly discussed the relationship between assessing writing ability and structuring the assignment. In his view, by providing students with the content in a writing test, one is filtering out, to some extent, the factor of subject matter mastery. On the other hand, when all of the content is provided, writing becomes simply an excercise in logic. He concluded that more experimentation is needed in this area in order to determine to what extent content should be provided in assessing writing ability and not knowledge of subject matter nor logic.

It is important to make a distinction between the use of the essay to assess ability to communicate within a subject area and the use of written compositions to assess ability to write. Coffman (1971) has addressed the former, but some of his ideas are relevant to the latter use. In particular, Coffman's chapter deals with the essay examinations when it is used by individual teachers in measuring the outcome of instruction.

In his chapter, Coffman considered three sources of error affecting essay scores: inter-rater variability, intra-rater variability, and freedom of responses. Not all three sources are pertinent to all uses of the essay. The last source is related to McColly's concern on the structure of the assignment. According to Coffman, if ratings are used only to determine the rank order of the pupils, only the first source of error is of concern. However, if the ratings are treated as direct measures of quality, then all sources of error become critical.

More recently,Cooper and Odell (1977) have noted that to obtain reliable measures of writing ability through essay tests, it is necessary to have more than one piece of writing from more than one occasion and involving two or more persons in rating each piece. Thus, these authors implied that raters, occasions, and assignment are sources of error.

A line of empirical studies addressing the issue of factors affecting specifically the writing of children clearly points out that writing mode is an important source of variation. Seegars, as early as 1933, cautioned teachers and researchers to be alert to the different impacts of the modes in evaluating and analyzing children's writing. Several experimental studies conducted in the 60's generally support Seegars' contention in samples of first and third grade children (Johnson, 1967; Anderson and Bashaw, 1968). More recent studies offer added evidence that the mode is related to the quality of children's writing (Bortz, 1970; Veal and Tillman, 1971; Pope, 1974; Perron, 1976).

In most of these studies, a measure of syntactic complexity such as number of clauses or number of words per clause was used as the dependent variable. The modes investigated were descriptive, argumentative, narrative, and expository.

In spite of the recognition that occasion variability, assignment variability, and mode variability are sources of error in assessing writing ability, most researchers who study compositional writing have considered the issue of instrument reliability in terms of inter-rater reliability. For example, Cohen (1973) in evaluating the writing ability of college students, determined reliability using percentage of agreement among raters. When Fagan, Cooper, and Jensen (1975) reviewed several available measures for evaluation and research in written language arts, inter-rater reliability or percentage of agreement between raters constituted the most common type of reliability estimates reported. The only other type of estimate, reported in only two cases, was test-retest reliability. More recent investigations of the reliability of specific instruments equate reliability with agreement across raters. An example is Singleton's (1977) dissertation on the reliability of ratings assigned on the essay portion of the Language Skills Examination.

It seems that essay test reliability has practically become synonymous with inter-rater reliability. A likely explanation for this phenomenon is that non-statistical psychologists find it easier to think in terms of correlations. A Pearson product-moment correlation coefficient may be easily computed between the scores assigned by two raters. But this correlation coefficient does not adequately assess all of the sources of variation (Coffman, 1971).

Coffman suggested using the analysis of variance approach to adequately assess more than one source of error variation. Stanley (1962) had previously discussed a specific design which could be used to assess the reliability of raters and test forms.

A classic study by Finlayson (1951) is the first reliability study to consider rater and test variability as sources of error in essays. Based on a sample of 197 children who wrote two essays, he reported mean coefficients of .697 and .810 for the reliability across tests and raters, respectively. Each essay was rated by six raters, using a general impression method of scoring with a 1 to 5 scale. In a second part to his study, Finlayson used the analysis of variance in a 197x2x6 random effects design. In testing the significance of effects he found the child-by-essay interaction significant, suggesting that the performance of a child in one essay is not representative of his/her ability to write in general. The child-by-rater interaction was not significant. From his results, it may be concluded that test variation represents a greater source of error than rater variation.

In a follow-up study, Vernon and Millican (1954) investigated the reliability across 7 raters and 7 topics for a sample of 224 college students using a general impression 5-point scale. They reported mean correlations between raters on the same topic and between topics. These were .509 and .366, respectively. In the authors words: "a still more serious source of inconsistency in assessing English ability is the varying performance of candidates when writing essays on different topics"(p.73).

In view of the recommendations made by language arts specialists and the results of the empirical studies reviewed, it appears that extending the design of Finlayson to include raters, modes/topic, and day-to-day variation as possible sources of error is in order. To best assess all sources simultaneously, the principles of generalizability theory will be applied. The development of generalizability theory will be discussed in the following section.

Generalizability Theory

The conceptual underpinnings of generalizability theory are based on Fisher's (1925) work on the analysis of variance, the factorial experiment, and the intraclass correlation.

The idea of using the analysis of variance to estimate the reliability of a test is due to Cyril Burt who translated the work of Fisher for his students with the aid of P. O. Johnson, J. Neyman, and R. W. B. Jackson (Burt, 1955). Burt considered measurements as varying in three dimensions; with respect to the person, the test form, and the occasion. The reliability of the test is estimable from a comparison of individual variance to group variance. In Burt's words:

> On comparing the two variances it would then seem possible, on intuitive grounds, to infer that, when the variance of the measurements for a single individual becomes as large as the variance for the entire sample of different individuals, the test used will be of no practical value whatsoever: for the whole object of such a test is to distinguish the ability as measured for any given individual from the abilities of the rest. (p.105)

Burt showed how the intraclass correlation provided an estimate of the reliability.

The intraclass correlation was introduced by Fisher in the context of the random effects model. Scheffe (1959) illustrates it using the model

$$(4) \qquad Y_{ij} = \mu + \alpha_i + e_{ij} \quad ,$$

where μ is the grand mean and α_i and e_{ij} are independent with zero means and variance matrices $\sigma^2(\alpha)I_i$ and $\sigma^2(e)I_i$, respectively. The variance of Y_{ij} may be expressed as

$$(5) \quad \sigma^2(y) = \sigma^2(\alpha) + \sigma^2(e) \ .$$

The observations within any class are not statistically independent. The statistical dependence between any two observations y_{ij} and $y_{ij'}$ in the same class is expressed as

$$\begin{aligned} (6) \quad \text{r intraclass} &= E[(y_{ij} - \mu)(y_{ij'} - \mu)] \ / \ \sigma^2(y) \\ &= E[(\alpha_i + e_{ij})(\alpha_i + e_{ij'})] \ / \ \sigma^2(y) \\ &= E(\alpha_i^2) \ / \sigma^2(y) \\ &= \sigma^2(\alpha) / [\sigma^2(\alpha) + \sigma^2(e)] \ . \end{aligned}$$

Thus, the intraclass correlation may be estimated by obtaining point estimates of the variance components.

Pilliner (1952) compared the estimate of reliability obtained from the intraclass correlation to that obtained from the Pearson product-moment correlation for a situation where measures vary in two dimensions: persons and tests (or items, etc.). Under homogeneity of variance assumptions, the intraclass correlation provides an unbiased estimate of reliability. But if variances are heterogeneous, the estimates from the intraclass r are negatively biased. That is, they represent a lower bound. Pilliner suggested extensions of the two dimensional framework where components of variance are mostly needed. His illustration was a three dimensional design using Finlayson's data, for which his procedures were derived.

In the United States, Hoyt (1941) used the analysis of variance approach in determining the internal consistency of a test from a subject-by-item design, where the items are dichotomously scored. He arrived at reliability formulas identical to those derived by Kuder and Richardson (1937).

Ebel (1951) made a case for the use of the intraclass correlation for situations where the parallelism assumption was impractical due to the inequality of means. He was interested in the reliability of ratings which he estimated by applying the analysis of variance to a subjects-by-ratings design. The results from this approach were compared to two other formulas proposed for estimating such reliability: the generalized reliability and the average intercorrelation. Ebel concluded that the intraclass formula was preferable because of its flexibility with respect to the inclusion of the between raters variance in the error term. In situations where the same raters are used to rate all subjects, the between raters variance does not enter into the error. On the other hand, when different raters are used, then that variance should be considered as error.

In his 1953 textbook, Lindquist provided a clear and comprehensive treatment of the use of variance components in the estimation of reliability. He discussed the possibility of obtaining negative estimates particularly when the number of degrees of freedom is small for some factors. A small number of degrees of freedom may not be crucial for variance components which are not of interest (such as the between raters variance discussed by Ebel in situations where all raters rate all subjects). Lindquist also demonstrated that increasing the number of observations in a study resulted in different effects, depending on the levels of the factors sampled. In this regard, the Spearman-Brown formula has limited utility. The limitations of the Spearman-Brown formula for showing the effects on reliability from an increase in the levels of a factor had been previously discussed by others (e.g., Pilliner). Finally, Lindquist illustrated the added utility of

estimating variance components for determining the relative importance of the various sources of error. This information could be useful in suggesting designs for the construction of measurement schedules. The idea of using variance component estimates for deciding among different designs was later expanded by Vaughn and Corballis (1969).

Using the analysis of variance approach and extending the designs used to estimate reliability to more than two dimensions implied a conceptualization of reliability as a characteristic of a measurement procedure rather than a measurement instrument. This was the position taken by Rajaratnam (1960) and, more recently, discussed by Rowley (1976) in the context of observational measures.

In his article, Rajaratnam introduced the notion of a reliability coefficient as the ratio of true score variance to the observed score variance expected in a set of observations obtained by using the same measurement procedure in a specific way. He formulated coefficients for situations where every rater does not rate every subject. In this situation, as Ebel had suggested, the systematic variance of raters is part of the error term since it enters into the expected observed score variance. Rajaratnam also introduced the distinction between G and D studies which was discussed in the introduction.

In studying the reliability of classroom observational schedules, Medley and Mitzel (1963) made use of the analysis of variance approach in reliability estimation. Their application is extended to a four-way factorial without replications. These authors illustrate the vast amount of reliability information which may be obtained from one carefully designed study using analysis of variance methods.

Several articles published by Cronbach, Gleser, and Rajaratnam (Cronbach et al., 1963; Gleser et al., 1965; Rajaratnam et al., 1965) and culminating in the publication of their 1972 book, have summarized the conceptualization of reliability estimation from the analysis of variance. These authors presented a general framework which encompasses the classical model and may be extended to include experimental designs for fixed, random, and mixed models. They rely heavily on the paper by Cornfield and Tukey (1956) dealing with variance component estimation for factorials through the use of expected mean squares. Their treatment is limited to balanced designs, having equal numbers of observations in the subclassifications.

In the introductory chapter, certain problems associated with the classical theory were presented. One of these problems was discussed by Guttman (1953) in his critique of Gulliksen's (1950) book. Guttman observed that the notion of parallel tests, the heart of classical reliability theory, does not provide a unique definition of reliability, since there may be more than one reasonable basis for forming parallel tests.

In their work, Cronbach et al. (1963) reformulated the theory of reliability to overcome the inadequacies presented by the parallelism assumption. They rephrased the reliability issue as follows: "an investigator asks about the precision or reliability of a measure because he wishes to generalize from the observation in hand to some class of observations to which it belongs"(p.144). Their theory requires that the investigator clearly specify a universe of conditions of observation over which generalization is to be made. The problem of reliability thus, becomes one of generalizability.

In terms of generalizability theory, a person's universe score, (analogous to the classical true score), is defined as the expected score over all admissible observations. This definition is equivalent to Lord and Novick's (1968) "generic true score." The obtained score is a sample from a universe of admissible observations and will generally differ from the universe score.

A model is constructed where the observed score is expressed in terms of the hypothesized effects. For example, consider the model

$$(7) \qquad X_{pj} = \pi_p + \alpha_j + e_{pj} \quad .$$

The observed score, X_{pj}, given to person p by judge j is the sum of three components, namely π_p , the effect for person p; α_j , the bias of judge j; and an error component e_{pj}, which may, for example, represent some idiosyncratic reaction of judge j to a particular person p. These components are assumed to be independent. Models like (7) can be constructed to fit any particular design.

The variation found among observed scores, $\sigma^2(X)$, may be partitioned into variance components

$$(8) \qquad \sigma^2(X) = \sigma^2(\pi) + \sigma^2(\alpha) + \sigma^2(e).$$

$\sigma^2(\pi)$ represents the variation due to persons and, in Cronbach's terms, the universe score variance.

Cronbach et al. (1972) make a distinction between two error components $\sigma^2(\Delta)$ and $\sigma^2(\delta)$. (this distinction was previously noted by Ebel (1951)). To illustrate the distinction assume that every judge considered every person. The component $\sigma^2(\delta)$, estimated from $\sigma^2(e)$ in our model, refers to the variance of each person's observed deviation scores under each judge, $(X_{pj} - \overline{X}_j)$, around the universe deviation score $(\mu_p - \mu)$. These deviation scores eliminate the systematic variance

among judges, $\sigma^2(\alpha)$, since the mean for each judge is subtracted from the raw score to obtain the deviation score. In general, the systematic variance of facets where the same conditions are sampled for every person is excluded from $\sigma^2(\delta)$. The error variance $\sigma^2(\Delta)$, refers to the variance of each person's observed scores, X_p, around their universe score, μ_p. In our example, $\sigma^2(\Delta) = \sigma^2(\alpha) + \sigma^2(e)$. In the classical sense, this variance component is the only component of error. The square root of $\sigma^2(\Delta)$ is the standard error of measurement. It will be noted that $\sigma^2(\Delta)$ will generally be greater than $\sigma^2(\delta)$.

The emphasis of generalizability theory is on the estimation of the variance components. These variance components have several uses, one of which is the estimation of generalizability coefficients via intraclass correlations. The coefficient of generalizability is defined as the ratio of the universe score variance to the expected observed score variance. It is approximately the expected value of the squared correlations of observed score and universe score, $E\rho^2(X_p\mu_p)$. The intraclass correlation is a good approximation of $\rho^2(X\mu)$ if homogeneity of variance assumptions are met. Maxwell and Pilliner (1968) and Selvage (1976) have recommended performing transformations on the data to achieve stability of variances when the assumptions are not met.

The variance components are also used in planning designs for D-studies. When making absolute decisions, it is desirable to reduce the error $\sigma^2(\Delta)$. According to Cronbach et al. (1972) a nested design reduces $\sigma^2(\Delta)$ more than a crossed design with the same number of observations per person since more conditions are sampled in the nested design. For comparative decisions $\sigma^2(\delta)$ is the appropriate error to consider in determining the adequacy of the measurement procedure. The magnitudes of the variance components provide an indication of the relative

contribution of the different effects to the error. This knowledge is useful in determining the number of conditions to be sampled from each facet in subsequent D studies in order to maintain the error at a specified level.

Cronbach et al. (1972) consider a third type of error, $\sigma(\varepsilon)$, the error of estimate. It is the square root of the familiar variance for errors of estimate in linear regression. The regression equation they consider is that for predicting μ_p from the observed score and group information. According to Cronbach et al. (1972, p.15), the universe score is "the ideal datum on which to base. . .decision [s]." They recommend estimating universe scores through linear regression and setting confidence intervals around the estimated true score using $\sigma(\Delta)$. The estimated universe scores are not very useful if all scores are regressed to the population mean; since they will be perfectly correlated with the observed scores. But if subpopulations of persons exist with different means, the universe score may be predicted from the observed score and the subpopulation information.

In their book, Cronbach et al. provide detailed examples of the application of generalizability theory to simple experimental designs involving both crossed and nested facets. They also extended the theory to encompass multivariate problems.

Since the publication of Cronbach's book several authors have applied the principles of generalizability theory to various situations. Levy (1974) applied the theory to studies of reliability in clinical settings; and Gillmore, Kane, and Naccarato (1978) to student ratings of instruction.

In the spirit of generality, Mellenbergh (1977) has recently proposed a more extended view of reliability by considering all possible replications of the design where in addition to replications of facets, replications of subjects for fixed facets is also possible. He suggested using replicability coefficients which are defined as the correlation between two replications of the design. His coefficients include generalizability coefficients and also make use of estimates of the variance components. Several of the possible coefficients, however, serve no interesting purpose in most practical situations.

Brennan (1975) extended the idea of calculating reliability from a person-by-item analysis of variance to a situation where persons are nested within some higher order dimension. Assuming an equal number of persons in each class, Brennan compared the generalizability coefficients derived from a split-plot factorial design with students nested within classes and crossed with items to those derived when the nesting classification (i.e. classes) is ignored (a randomized blocks design). He concluded that "the experimental model used to collect data for most reliability studies is usually one where students are nested within some dimension; therefore, the split-plot design would appear to be more appropriate than a simple randomized block design" (p.780). In addition, the split-plot design can be used to provide a basis for estimating the reliability of scores for the units within which persons are nested.

For his design, Brennan stated that the reliability of the test of specified length calculated from the split-plot design would be less than, equal to, or greater than that calculated from a randomized block design depending upon whether the ratio of $\sigma^2(\rho)$ (the person

variance component) to $\sigma^2(e)$ (the error variance) is less than, equal to, or greater than the ratio of $\sigma^2(s)$ (the school variance component) to $\sigma^2(si)$ (the school by item variance component).

> Thus, if one uses a randomized block design to calculate reliability for persons when, in fact, persons are nested within some dimension, such as schools or classrooms, the resulting coefficient will be biased, and, moreover, the direction of bias will be unknown. (p.785)

Kane and Brennan (1977) extended generalizability theory to a split-plot design in which students were nested within classes and crossed with items. Their purpose was to estimate the generalizability of a class mean, where the class was the unit of analysis. They assumed an equal number of students in each class. Four different coefficients were formulated corresponding to four universes: an infinite universe of students and items, a fixed universe of students and items, a universe with fixed students and infinite items, and a universe with infinite students and fixed items.

The situation where the students are fixed is somewhat artificial since, in educational research, it is generally inappropriate to restrict the universe of generalization for the student facet. Restricting the set of both items and students is very unlikely. The universe score variance in this case is estimable if the interaction effect for students and items and the error in the model are not confounded, that is, if there is more than one replication of each class-student-item observation or if the student-item interaction is assumed to be zero and its estimate taken as the error estimate.

In a subsequent section, the authors showed how certain coefficients may be estimated from mixed models. However, since the components

from a model with a fixed facet cannot be used to estimate a generalizability coefficient that assumes generalization over that facet, the authors recommended a random model in the estimation of variance components.

Kane and Brennan also related three coefficients, which appear in the literature for estimating the reliability of class means, to their four generalizability coefficients. None of the four reliability coefficients is equivalent to their generalizability coefficient where generalization is intended over students and items, a very common situation.

Generalizability theory offers innumerable possibilities for well designed studies to be conducted as part of instrument development. Much information may be gained from one G-study, some of which is unattainable under the classical approach. As the principles are applied to various measurement problems, their strengths and limitations will become apparent.More applications are needed in all areas. To this author's knowledge the theory has not been applied to the assessment of writing ability. The studies by Finlayson (1951) and Vernon and Millican (1954) approximate this effort. However, these studies only reported tests of hypotheses and interclass correlation coefficients and did not use estimates of variance components. This applied study extended the design used by Finlayson and incorporated a method of estimating variance components for unbalanced data.

Variance Component Estimation

Thus far, all references to generalizability theory,both theoretical and applied,have assumed balanced designs. For balanced designs the analysis of variance method of estimation is universally accepted. The

expected values of the mean squares may be expressed as linear combinations of the variance components. The coefficients of the components are easy to obtain by rules developed by Cornfield and Tukey (1956) for fixed, random, and mixed models. These rules appear in standard texts such as Kirk (1968) and Winer (1971). The best method of estimation is to equate the observed mean squares from the analysis of variance under fixed effects, to the linear combination of variance components. Then the resulting set of simultaneous equations is solved for the variance components. These estimates are minimum variance and are unbiased (Searle, 1971b).

Most methods of estimating variance components involve some quadratic form of the observations. The mean squares from the analysis of variance are the appropriate quadratics to use when the design is balanced. Estimating variance components from unbalanced data is more complex because there is no universally accepted method. According to Searle (1971b, p.33) "no particular set of quadratics has been established as being more optimal than any other set." For unbalanced designs, using the analysis of variance procedure leads to the question of which mean squares to use, since with unbalanced data the mean squares may be unadjusted or adjusted for one or more effects.

A comprehensive review of methods of estimation based on the analysis of variance has been given by Searle (1971a, 1971b) for both balanced and unbalanced designs. For the latter case, Searle discussed three methods proposed by Henderson (1953). Henderson's method 1 consists of equating the unadjusted sums of squares from the fixed effects analysis of variance to their expectations obtained under a random

effects model. These expectations are linear combinations of the variance components. Thus, solving for the set of simultaneous equations will yield estimates of the components. This method produces unbiased estimates except for the random effects in mixed models.

Henderson's method 2 was developed to correct the inefficiency of method 1 with mixed models. The procedure of the second method is to "correct" the data by some previous least squares estimates of the fixed effects. Using the "corrected" data in place of the original data, method 2 proceeds as method 1. This method is inappropriate when there are interactions between the fixed and random effects.

The method of fitting constants, or Henderson's method 3, uses the adjusted sums of squares--adjusted sequentially--and follows the same pattern as the other methods. The adjusted sums of squares are similar to those of Overall and Spiegel's (1969) "a priori ordering." All expectations of these adjusted sums of squares are taken under the full model. Under this condition, the expected value of any term involves all of the variance components except those for the terms for which this term was adjusted.

As Searle pointed out, the coefficients of the variance components for these methods are not as easy to obtain as those with balanced data. He gives several references which discuss numeric methods for obtaining the coefficients.

More recently Rao (1971, 1972) has proposed a different approach to the estimation of variance components. His methods called MINQUE (minimum norm quadratic unbiased estimation) and MIVQUE (minimum variance quadratic unbiased estimation) provide a general approach which is

applicable to both balanced and unbalanced designs and suitable for either random or mixed models.

To summarize them, let us consider the model

$$(9) \quad \underline{Y} = \underline{XB} + \underline{U}_1 \underline{\xi}_1 + \underline{U}_2 \underline{\xi}_2 + \ldots + \underline{U}_k \underline{\xi}_k \quad .$$

where $\underline{Y}$ is the n x 1 vector of observations, $\underline{X}$ is a n x m design matrix for the fixed effects (in a random effects model $\underline{X}$ is just a column vector of 1's), $\underline{B}$ is a vector of unknown parameters (the grand mean in a random effects model), $\underline{U}_i$ is a given n x c_i matrix, the columns of which are the coded variables for a particular factor, and $\underline{\xi}_i$ is a c_i vector of uncorrelated variables for the ith random effects factor in the model (which may be a main effects factor or an interaction factor). The $\underline{\xi}_i$'s have zero mean and variance matrix $\sigma_i^2 \ \underline{I}_{ci}$, i=1,. . ., k, where σ_i^2 are unknown. Furthermore, $\underline{\xi}_i$ and $\underline{\xi}_j$ $(i \neq j)$ are uncorrelated. The kth factor is the error term. Then

$$(10) \quad E(\underline{Y}) = \underline{X}\ \underline{B},$$

$$(11) \quad \underline{V}^* = \mathrm{Var}(\underline{Y}) = \sigma_1^2 \underline{V}_1 + \ldots + \sigma_k^2 \underline{V}_k \quad \text{where}$$

$$(12) \quad \underline{V}_i = \underline{U}_i \underline{U}_i' \quad .$$

Rao defined

$$(13) \quad \underline{V} = \sum_{i=1}^{k} \underline{V}_i \quad .$$

The problem then is to estimate the variance components $\sigma^2_1, \ldots, \sigma^2_k$.

Rao considered the estimation of a linear function

$$(14) \quad p_1\sigma_1^2 + \ldots + p_k\sigma_k^2$$

of the variance components from a quadratic function $\underline{Y}'\ \underline{A}\ \underline{Y}$ of the observations.

$\underline{A}$ is symmetric and is chosen to satisfy the following conditions:

$$(a) \quad \underline{A}\ \underline{X} = \underline{0}$$

$$(b) \quad E(\underline{Y}'\underline{AY}) = p_1\sigma_1^2 + \ldots + p_k\sigma_k^2 \quad .$$

Condition (a) is necessary for the estimator to be invariant to changes in $\underline{B}$ (Rao, 1972). For condition (b) to be true (i.e. the estimator is unbiased), then tr $\underline{A}\,\underline{V}_i = p_i$, i = 1, . . ., k where tr represents the trace of a matrix (the sum of the diagonal elements). To obtain the MINQUE estimator, the Euclidean norm tr $(\underline{V}^*\underline{A})^2$ is minimized. This requires some a priori knowledge of the ratios of σ^2_i. To obtain the MIVQUE estimator, the variance of $\underline{Y}'\underline{A}\,\underline{Y}$ is minimized for a particular choice of $\sigma_1, \ldots, \sigma_k$. That variance is Var $(\underline{Y}'\underline{A}\,\underline{Y}) = 2\,\text{tr}(\underline{V}^*\underline{A})^2$ + a term in $\underline{A}$ and kurtosis parameters. Under normality assumptions, the kurtosis parameters are zero and MINQUE equals MIVQUE.

Rao's methods are preferable to those proposed by Henderson for three reasons. First, they have a wider range of applicability since they can accommodate mixed as well as random models. Second, the computations involved are more efficiently programmable. Third, when prior estimates of the components are available, the MIVQUE method provides estimates which are locally minimum variance.

The second reason is relevant to G-studies because the designs used in such studies tend to be large. As was mentioned previously, a G-study should include as many sources of error variance related to a measurement procedure as possible. For each facet included, the maximum number of conditions possible should be sampled. The resulting design then requires the most efficient method for its analysis. Rao's methods satisfy this criterion.

Summary

The literature pertinent to the measurement of writing ability indicates that essay tests represent the most valid method of assessment. Several sources of error have been identified as affecting the

reliability of this test form. Although variability among raters is the source most commonly examined, day-to-day and assignment variability are considered to be equally or more important. Missing from the literature are empirical studies which examine how these sources affect the reliability of the essay.

Generalizability theory offers a conceptual framework which is applicable to the study of multiple sources of error variation. Based on Fisher's work on the analysis of variance, in the theory, the problem of reliability is considered as one of generalization from one observation to a universe of admissible observations.

In a generalizability study, the observations are gathered under a specific design characterized by facets, the identified sources of error. The conditions of each facet included in the design may be fixed or sampled from the total universe of conditions. The relative magnitude of the sources of error variation is determined through the estimation of variance components. For purposes of simplification in the estimation process, generalizability theory has been restricted to balanced designs. The literature on generalizability theory is lacking in applied studies, although content areas such as writing could greatly profit from its application.

Also missing from the psychometric literature are extensions of the theory to unbalanced designs. These extensions are much needed since these designs are typical in educational research. Psychometricians could profit from methods of estimating variance components documented in the statistical literature. In particular, the methods of Henderson (1950) and Rao (1971, 1972) are applicable to unbalanced data. These methods allow the principles of generalizability theory to be further extended.

CHAPTER III

METHOD

This study was designed to demonstrate the application of generalizability theory to the assessment of writing ability. The data were collected in a natural setting on a sample of fourth grade children. The study extended the application of the theory to a situation where unequal but proportional numbers of subjects appeared in the subclassifications.

The sample, the facets, the design, and the procedures for data collection and analysis are described in this chapter.

The Sample

The sample used in this study consisted of 104 fourth grade students from eight classes in two schools in Alachua county; four from P. K. Yonge Laboratory School and four from Alachua Elementary School. The data used in this study were collected as part of a research project on creative writing conducted at those schools.

P. K. Yonge is a laboratory school associated with the College of Education at the University of Florida. The student population at each grade level is selected from a waiting list in such a way to approximate, in each classroom, an equal balance between males and females; a 20:80 racial balance between blacks and white or others, respectively; and an equal balance from each of five income categories. Fourth and fifth grades are combined in the classrooms at this school.

The four classrooms participating in this study exhausted those classrooms containing fourth grade students. A total of 59 fourth grade students are currently enrolled at P. K. Yonge. However, only 37 who had complete data were used in this study.

Alachua Elementary is a public school in the rural town of Alachua. The four classrooms from this school also exhausted the fourth grade population. In this school, students at each grade level are assigned to classrooms to maintain the sex and race balance previously described. A total of 67 students in this school had complete data out of an initial sample of 113. Thus, the writing samples used in this study were obtained from a total of 104 individuals. The sample sizes for each class are shown in Table 1, broken down by sex and race.

The Writing Samples: Data Collection

Samples of compositional writing, in two different writing modes, were collected on three occasions. On each occasion, verbal and written instructions were given to the children by one of the staff members of the project. The same person collected the samples throughout the occasions at each school. Steps were taken to insure that the children understood the task. Furthermore, praise was used in an attempt to motivate the children to write. On each occasion, the assignment and the instructions were standard for all students. Each student was allowed sufficient time to complete the task. On the average, the compositions were completed in approximately 45 minutes.

The Facets

The writing samples were characterized by two facets: modes and occasion. A third facet, raters, was introduced in scoring the samples. The levels of these facets which were used in this study are described next.

TABLE 1

SAMPLE SIZES BY CLASSROOM, SEX, AND RACE

Classroom	Sex		Race		Total
	Male	Female	Black	White	
1	3	5	1	7	8
2	4	4	2	6	8
3	8	3	2	9	11
4	3	7	3	7	10
5	6	11	4	13	17
6	8	11	7	12	19
7	6	5	2	9	11
8	8	12	4	16	20

Note: Classrooms 1 through 4 are from P. K. Yonge and 5 through 8 are from Alachua Elementary.

Modes

The mode facet, as conceptualized in this study, was characterized by two dimensions: the purpose of the writing sample and the type of assignment. This use of the word mode is broader than the traditional use. Generally, four basic writing modes are mentioned in the literature related to factors which influence children's writing ability. These are: narrative, declarative, argumentative, and expository. Each of these modes constitutes a different purpose. For example, the purpose of writing in the narrative mode is to tell a story; that of the argumentative mode is to convince the audience. For each one of these purposes, different types of assignments are possible. A child who is asked to write in the narrative mode may tell his/her story through a poem, a letter, a report, etc. Characterizing the type of writing along these two dimensions allows for a large number of possible conditions on this facet. In this study, generalization was intended to all of the possible conditions thus identified.

Two types of writing assignment were used in this study, each representing a different writing purpose. In one mode, children were instructed to prepare a brief report about specific animals using a standard set of facts supplied by the investigator. The facts were presented either in written form or with the aid of a film. On the first occasion a list of facts about bats was provided for the children. Films about cows and pigs provided the facts used on the second and third occasions, respectively. After the presentation of the stimuli, the facts were discussed with the children.

In the second mode, the children were asked to write a creative story explaining some imaginary phenomemon such as "how the camel got

the hump". On each occasion, a list of titles was provided for the children from which they were to select one.

Occasions

The writing samples were collected on three occasions during the 1977-78 school year: Fall, Winter, and Spring. On each occasion, the descriptive reports were collected one week before the narrative stories. This order was maintained because the investigators felt that there would be less carry-over from a report to a story than vice versa. The one week time period within an occasion was allowed for two reasons: to minimize carry-over effects and to maximize the motivation of the children. With children at the elementary level, there is a loss in motivation when similar tasks are assigned in the same day.

Writing performance is expected to fluctuate from day to day. Furthermore, it is expected that children's writing ability will also fluctuate (hopefully improve) during the year. In this study, generalization was intended to any time during the school year.

Raters

The four raters represent a sample of raters which could have been used. Three of the raters were graduate students in educational research; the fourth rater, an associate professor in the same department. Generalization along this facet is intended to any person who would rate a sample of writing for the purpose of making a decision about placement, selection, grading, or for purposes of comparison in a research study.

The writing samples were collected and sorted into six modes-by-occasion combinations. The children's names were covered and a number was assigned and written on their sample for identification. Thus, the anonymity of the samples was preserved. The raters scored the

samples on eight different days. Each day, the four raters scored the samples using a general impression scoring method. At the beginning of each scoring session, the raters reviewed the criteria to be used in scoring. After scoring several samples, the raters compared their scores and discussed samples which had received divergent scores. These discussions were an attempt to increase the inter-rater reliability. Each sample was scored independently.

Prior to the first rating session, the raters were trained in using the general impression method. Samples from fifth grade students were used for training. During training, the scaling points were determined so as to obtain an approximation to a normal distribution. Normality was not a consideration during the actual scoring of the samples. A general impression method of scoring used in this study involved assigning a score of 1 through 8 on the basis of the overall quality of the writing sample. The method involves the rapid, impressionistic scoring of a sample. Generally, no more than two minutes are spent on any one paper.

This procedure has been used by the Educational Testing Service (ETS) and the College Entrance Examination Board, and was also used in the first national assessment of writing conducted by the National Assessment of Educational Progress (NAEP) (Mellon, 1975). The ETS research on rater reliability in the 1960's revealed that multiple ratings based on overall impressions were the best means of achieving inter-rater reliability (Suhor, 1977). An additional advantage to this method is the fact that it requires less time than any other method.

Design

A schematic representation of the design used in this study is

FIGURE 1

SCHEMATIC REPRESENTATION OF THE
DESIGN INCLUDING CLASSES(C), STUDENTS(S),
OCCASIONS(O), MODES(M), AND RATERS(R)

		O_1		O_2		O_3	
		M_1	M_2	M_1	M_2	M_1	M_2
		$R_1R_2R_3R_4$	$R_1R_2R_3R_4$	$R_1R_2R_3R_4$	$R_1R_2R_3R_4$	$R_1R_2R_3R_4$	$R_1R_2R_3R_4$
C_1	S_1						
	S_8						
C_2	S_9						
	S_{16}						
C_3	S_{17}						
	S_{27}						
C_4	S_{28}						
	S_{37}						
C_5	S_{38}						
	S_{54}						
C_6	S_{55}						
	S_{73}						
C_7	S_{74}						
	S_{84}						
C_8	S_{85}						
	S_{104}						

shown in Figure 1. The design is referred to as a split-plot factorial design in standard texts (e.g., Kirk, 1968 ; Winer, 1971) with the classes being the main plots and the students being the subplots. In this design students are nested within the classes; that is, each student appears in only one class. This situation is typical in the natural setting. The nesting of students within class results in the confounding of the student by class interaction with the student effect. As a result, there is no way to estimate the student effect independent of the class-by-student interaction. Similarly, any interaction term involving the student effect is confounded with the corresponding interaction term involving class-by-student. Since typically students are nested within classrooms, the confounding of the effects mentioned above does not present any problems.

The students, modes, occasions, and raters are factorially combined. In terms of this study, this factorial combination means that each student was measured in both modes on each occasion and that each rater scored every writing sample. The crossing of students, modes, occasions, and raters allows for the independent estimation of each main effect and all interactions involving those effects. The levels of all factors included in this study were considered to be random samples of all possible levels which could have been included. Thus, the model used is a random effects model.

Let X_{csmor} denote the rating received by students in class c for mode m , occasion o , and rater r . Then the structural model used in this study may be represented as:

$$(15)\quad X_{csmor} = \mu + \alpha_c + \pi_{s(c)} + B_m + \alpha B_{cm} + B\pi_{ms(c)} + \gamma_o + \alpha\gamma_{co} + \gamma\pi_{os(c)} + \theta_r + \alpha\theta_{cr} + \theta\pi_{rs(c)} + B\gamma_{mo} + \alpha B\gamma_{cmo} + B\gamma\pi_{mos(c)} + B\theta_{mr} + \alpha B\theta_{cmr} + B\theta\pi_{mrs(c)} + \gamma\theta_{or} + \alpha\gamma\theta_{cor} + \gamma\theta\pi_{ors(c)} + B\gamma\theta_{mor} + \alpha B\gamma\theta_{cmor} + B\gamma\theta\pi_{mors(c)}$$

where μ = the grand mean,

α_c = the effect for class c ($c = 1, \ldots, n_c$).

$\pi_{s(c)}$ = the effect for student s nested within class c ($s = 1, \ldots, n_{s(c)}$),

B_m = the effect for mode m ($m = 1, \ldots, n_m$),

αB_{cm} = the class-by-mode interaction effect,

$B\pi_{ms(c)}$ = the mode-by-student (nested within class c) interaction effect,

γ_o = the effect for occasion o ($o = 1, \ldots, n_o$),

$\alpha\gamma_{co}$ = the class-by-occasion interaction effect,

$\gamma\pi_{os(c)}$ = the occasion-by-student (nested within class c) interaction effect,

θ_r = the effect for rater r ($r = 1, \ldots, n_r$),

$\alpha\theta_{cr}$ = the class-by-rater interaction effect,

$\theta\pi_{rs(c)}$ = the rater-by-student (nested within class c) interaction effect,

$B\gamma_{mo}$ = the mode-by-occasion interaction effect,

$\alpha B\gamma_{cmo}$ = the class-by-mode-by-occasion interaction effect,

$B\gamma\pi_{mos(c)}$ = the mode-by-occasion-by-student (nested in class c) interaction effect,

$B\theta_{mr}$ = the mode-by-rater interaction effect,

$\alpha B\theta_{cmr}$ = the class-by-mode-by-rater interaction effect,

$B\theta\pi_{mrs(c)}$ = the mode-by-rater-by-student (nested in class c) interaction effect,

$\gamma\theta_{or}$ = the occasion-by-rater interaction effect,

$\alpha\gamma\theta_{cor}$ = the class-by-occasion-by-rater interaction effect,

$\gamma\theta\pi_{ors(c)}$ = the occasion-by-rater-by-student (nested in class c) interaction effect,

$B\gamma\theta_{mor}$ = the mode-by-occasion-by-rater interaction effect,

$\alpha B\gamma\theta_{cmor}$ = the class-by-mode-by-occasion-by-rater interaction effect, and

$B\gamma\theta\pi_{mors(c)}$ = the mode-by-occasion-by-rater-by-student (nested in class c) interaction effect.

It is assumed that each effect in the model (except for the grand mean) is a random variable with a mean of zero and variance σ^2(effect). The effects are assumed to be independent of each other so that the total variance in the scores X_{csmor} can be partitioned as

(16)

$$\sigma^2(x) = \sigma^2(\alpha) + \sigma^2(\pi) + \sigma^2(B) + \sigma^2(\alpha B) + \sigma^2(B\pi) + \sigma^2(\gamma) + \sigma^2(\alpha\gamma) + \sigma^2(\gamma\pi) + \sigma^2(\theta) + \sigma^2(\alpha\theta) + \sigma^2(\theta\pi) + \sigma^2(B\gamma) + \sigma^2(\alpha B\gamma) + \sigma^2(B\gamma\pi) + \sigma^2(B\theta) + \sigma^2(\alpha B\theta) + \sigma^2(B\theta\pi) + \sigma^2(\gamma\theta) + \sigma^2(\alpha\gamma\theta) + \sigma^2(\gamma\theta\pi) + \sigma^2(B\gamma\theta) + \sigma^2(\alpha B\gamma\theta) + \sigma^2(B\gamma\theta\pi).$$

The variances $\sigma^2_{\alpha}, \ldots, \sigma^2_{B\gamma\theta\pi}$ are called variance components (Scheffé, 1959) and, therefore, the model is referred to as a variance component model. To estimate variance components it is not necessary to assume that the effects are normally distributed.

Variance Component Estimation

To estimate the variance components in (16), a new version of the SAS VARCOMP procedure was used (Goodnight, 1978). This procedure, called MIVQUE0, is based on the MIVQUE (minimum variance quadratic estimator) method developed by Rao (1971). The method estimates linear functions of the variance components through the use of quadratic functions of the observations which have minimum variance for a particular choice of $\sigma_1, \ldots, \sigma_k$. The VARCOMP program selects $\sigma_1, \ldots, \sigma_k$ so as to minimize the ratio of the variance for each effect to the residual variance. The resulting estimates are invariant, locally best (at zero) quadratic unbiased estimates of the variance components (Goodnight, 1978). The program used was the only one available to handle the size of the design matrix within a reasonable amount of computer time and space.

For balanced split-plot factorial designs the expected mean squares are linear combinations of the variance components. In this case, the observed mean squares from the analysis of variance may be used in the formulas shown in Appendix A to estimate the variance components.

Generalizability Coefficients

Tests for homogeneity of variances were performed on the basis of warnings by Cronbach et al. (1972, p. 100-101). In their words:

> Where there is crossing of persons with facet i (or j, etc.) observed-score variances may differ from one application of the design to the next, and intercorrelations between pairs of independently obtained observed scores may differ. The intraclass correlation (our coefficient of generalizability) truly equals the mean of $\rho^2(X,\mu_p)$

only if all observed-score variances are equal. One must be hesitant, then, in taking the coefficient of generalizability as representing the parameter $\rho^2(X,\mu_p)$ for any particular D-study with crossed conditions.

To test for violations of homogeneity assumptions for this design, a procedure suggested by Box (1950) and recommended by Kirk (1968) was used. The procedure involved the following:

1. testing the equality of the variance-covariance matrices across the eight classes; and if this hypothesis was not rejected,

2. testing the equality of the diagonal elements in the pooled variance-covariance matrix.

The first test is performed by the DISCRIM procedure in SAS (Barr, et al., 1976). The second test was done using Bartlett's test for homogeneity of variance. It was recognized that this test is sensitive to violations of normality assumptions. However, a visual inspection of the frequencies within each subclassification revealed no serious departure from normality.

Using the point estimates of the variance components, one can derive the formulas for any desired coefficient of generalizability, where generalization is intended to any subset of the universe of generalization used in this study. The generalizability coefficient, $\rho^2(X,\mu)$, is defined as the ratio of $\sigma^2(\mu)$, the universe score variance, to $E(\sigma^2(X))$, the expected value of the observed score variance, the expectation taken over repeated applications of this design.

The universe of generalization determines what constitutes the universe score variance and the expected observed score variance. The expected observed score variance is always made up of the universe score variance plus error variance (Cronbach's $\sigma^2(\delta)$).

For deviation scores, the expected observed score variance is

(17)

$$E(\sigma^2(x)) = \sigma^2(\pi) + \frac{\sigma^2(\pi B)}{n_m} + \frac{\sigma^2(\pi\gamma)}{n_o} + \frac{\sigma^2(\theta\pi)}{n_r} + \frac{\sigma^2(B\gamma\pi)}{n_m n_o} + \frac{\sigma^2(B\theta\pi)}{n_m n_r} + \frac{\sigma^2(\gamma\theta\pi)}{n_o n_r} \quad \frac{\sigma^2(B\gamma\theta\pi)}{n_m r_o n_r} \quad .$$

It includes all of the components of variance involving the student effect. Other components in the model do not enter into the expected observed score variance because they are constant for all students, and in the formula, the students are considered in relation to the group's universe score, Each component is divided by the number of conditions entering the facet involved on that component. Given the formula for the expected observed score variance, the universe score variance may be obtained by taking the limit of $E(\sigma^2(X))$ as the number of conditions approaches infinity. This is the case when generalization is intended to an infinite number of levels, where all terms but $\sigma^2(\pi)$ disappear from the formula. Thus, $\sigma^2(\pi)$ is the universe score variance. In the situation where generalization is intended to a fixed number of conditions for a particular facet, the component involving that facet is considered as part of the universe score variance.

In Table 2, seven coefficients are suggested for all possible combinations of fixed and infinite generalizations across the three facets. These formulas may be used in a D-study involving a similar population of subjects and a subset of these facets by substituting the values for the n's for that study and these estimates of the variance components. The denominator of the formulas is the same for all universes. The terms have been rearranged so that the error component is within the parenthesis.

The Error Variance $\sigma^2(\Delta)$

The coefficients of generalizability exclude systematic facet components from the error term and, therefore, from the expected

TABLE 2

FORMULAS FOR GENERALIZABILITY COEFFICIENTS FOR THE SPLIT-PLOT FACTORIAL DESIGN WITH THREE FACETS RATERS(R), MODES(M), AND OCCASIONS(O) AND SEVEN UNIVERSES OF GENERALIZATION

Universe of Generalization	Formula
R infinite M infinite O infinite	$\dfrac{\sigma^2(\pi)}{\sigma^2(\pi) + \left[\dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$
R fixed M infinite O infinite	$\dfrac{\sigma^2(\pi) + \dfrac{\sigma^2(\pi\theta)}{n_r}}{\sigma^2(\pi) + \dfrac{\sigma^2(\pi\theta)}{n_r} + \left[\dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$
R infinite M fixed O infinite	$\dfrac{\sigma^2(\pi) + \dfrac{\sigma^2(\pi B)}{n_m}}{\sigma^2(\pi) + \dfrac{\sigma^2(\pi B)}{n_m} + \left[\dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$
R infinite M infinite O fixed	$\dfrac{\sigma^2(\pi) + \dfrac{\sigma^2(\pi\gamma)}{n_o}}{\sigma^2(\pi) + \dfrac{\sigma^2(\pi\gamma)}{n_o} + \left[\dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$
R fixed M fixed O infinite	$\dfrac{\sigma^2(\pi) + \dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r}}{\sigma^2(\pi) + \dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \left[\dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$
R fixed M infinite O fixed	$\dfrac{\sigma^2(\pi) + \dfrac{\sigma^2(\pi\gamma)}{n_o} \quad \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r}}{\sigma^2(\pi) + \dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \left[\dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$
R infinite M fixed O fixed	$\dfrac{\sigma^2(\pi) + \dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o}}{\sigma^2(\pi) + \dfrac{\sigma^2(\pi B)}{n_m} + \dfrac{\sigma^2(\pi\gamma)}{n_o} + \dfrac{\sigma^2(\pi B\gamma)}{n_m n_o} + \left[\dfrac{\sigma^2(\pi\theta)}{n_r} + \dfrac{\sigma^2(\pi B\theta)}{n_m n_r} + \dfrac{\sigma^2(\pi\gamma\theta)}{n_o n_r} + \dfrac{\sigma^2(\pi B\gamma\theta)}{n_o n_m n_r}\right]}$

observed score variance. Two situations may arise where these variance components should be considered as part of the error component. These are:

1. Studies where the conditions of a facet are nested within the student, rather than crossed. In other words, a different condition or set of conditions is sampled for each student.

2. Situations which involve determining confidence intervals around as individual's score for the purpose of making an absolute decision.

The formulas for estimating $\sigma^2(\Delta)$ from this design for different universes of generalization may be obtained from the information in Table 3. The entries in the table indicate those components which enter into the error variance. The components are to be divided by the frequencies shown in the last column of the table.

Summary

A total of 104 fourth grade students in eight classes participated in this study. Samples of compositional writing, in two different writing modes, were collected on three occasions. The samples were scored by four trained raters using an 8-point general impression method.

The design used, a split-plot factorial, considered the students as nested in the classes and crossed with the raters, modes, and occasions. A model was constructed which expressed the variance among all observations as a linear combination of independent variance components.

Estimates of the variance components in the model were obtained using the MIVQUEO method in SAS. This procedure is applicable to unbalanced designs such as the one considered in this study. Prior to

TABLE 3

VARIANCE COMPONENTS ENTERING INTO THE ERROR VARIANCE $\sigma^2(\Delta)$
FOR SEVEN UNIVERSES OF GENERALIZATION

	UNIVERSE OF GENERALIZATION							
Variance Component	R Infinite M Infinite O Infinite	R Fixed M Infinite O Infinite	R Infinite M Fixed O Infinite	R Infinite M Infinite O Fixed	R Fixed M Fixed O Infinite	R Fixed M Infinite O Fixed	R Infinite M Fixed O Fixed	Number of Replications Within π
$\sigma^2(B)$	*	*		*		*		n_m
$\sigma^2(\pi B)$	*	*		*		*		n_m
$\sigma^2(\gamma)$	*	*	*		*			n_o
$\sigma^2(\pi\gamma)$	*	*	*		*			n_o
$\sigma^2(\theta)$	*		*	*			*	n_r
$\sigma^2(\pi\theta)$	*		*	*			*	n_r
$\sigma^2(B\gamma)$	*	*	*	*	*	*		$n_m n_o$
$\sigma^2(\pi B\gamma)$	*	*	*	*	*	*		$n_m n_o$
$\sigma^2(B\theta)$	*	*	*	*		*	*	$n_m n_r$
$\sigma^2(\pi B\theta)$	*	*	*	*		*	*	$n_m n_r$
$\sigma^2(\gamma\theta)$	*	*	*	*	*		*	$n_o n_r$
$\sigma^2(\pi\gamma\theta)$	*	*	*	*	*		*	$n_o n_r$
$\sigma^2(B\gamma\theta)$	*	*	*	*	*	*	*	$n_m n_o n_r$
$\sigma^2(\pi B\gamma\theta)$	*	*	*	*	*	*	*	$n_m n_o n_r$

Note: the asterisks indicate those components which enter into the error variance.

using the estimates of the variance components in the estimation of generalizability coefficients, tests for homogeneity of variance were performed.

Formulas for generalizability coefficients corresponding to seven universes of generalization were provided. In addition, components of variance entering the formulas for the standard error of measurement were listed for seven universes of generalization. These universes represented generalization across one dimension (raters, modes, or occasions), two dimensions (raters and modes, etc.), or three dimensions (raters, modes, and occasions).

CHAPTER IV

RESULTS

This study was designed to apply the principles of generalizability theory to the assessment of writing ability in young children. Samples of writing from fourth grade children were collected in two modes at each of three occasions during the school year. A general impression method of scoring was used by four trained raters.

Because the children were nested in the classes, the observations were first considered in a split-plot factorial design with unequal numbers of subjects in the classes. The variance components for all effects in this model were estimated using the MIVQUE method.

A model ignoring the class dimension was also considered. For this second model, estimates of the variance components were obtained through the analysis of variance mean squares. The results from these methods are reported in this chapter. Also reported here are the results of the homogeneity of variance tests as well as certain coefficients of generalizability and error variances, $\sigma^2(\Delta)$.

Estimates of the Variance Components

The point estimates of the variance components in model (16), obtained from the MIVQUE0 method of SAS are reported in Table 4 along with their corresponding degrees of freedom. Negative estimates were replaced by zeros, following the recommendation of Cronbach et al. (1972) among others. These zero estimates are no longer unbiased

(Searle, 1971b, p.23) and are obviously bad estimates since a variance is, by definition, non-negative.

Searle (1971b) suggested six courses of action to follow when negative estimates of variance components are obtained. Three of these alternatives involve assuming that the true value is zero. The first one is to report the negative estimate but use it as evidence that the true value is zero. The second one is to change the negative estimate to zero, as was done in this study. The third involves ignoring the negative components from the model and reestimating the other components. The fourth is to use the negative estimate as an indication of an inappropriate model for the data and to reconsider the model, possibly considering models with finite instead of infinite populations. The fifth course of action is to use Bayesian or maximum likelihood estimators. The last recommendation suggested by Searle is "the statistician's last hope", to collect more data.

As shown in Table 4, seven out of the 23 estimates are considered to be zero. The actual estimates were very small. In general, all estimates of the variance components were small. This may be partially due to the restricted range imposed by the 1 to 8 rating scale.

The largest estimates were for the student effect ($\hat{\sigma}^2(\pi) = .346$), the student-by-mode-by-occasion interaction ($\hat{\sigma}^2(B\gamma\pi) = .339$), and the student-by-mode-by-occasion-by-rater interaction which is confounded with the error ($\hat{\sigma}^2(B\gamma\theta\pi) = .235$). Following in order of magnitude were the student by occasion interaction ($\hat{\sigma}^2(\gamma\pi) = .073$) and the occasion main effect ($\hat{\sigma}^2(\gamma) = .070$). All other estimates appear negligible.

TABLE 4

POINT ESTIMATES OF THE VARIANCE COMPONENTS
FOR THE MODEL (16)

VARIANCE COMPONENT	df	POINT ESTIMATE
$\hat{\sigma}^2(\alpha)$	7	0.000*
$\hat{\sigma}^2(\pi)$	96	0.346
$\hat{\sigma}^2(B)$	1	0.000*
$\hat{\sigma}^2(\alpha B)$	7	0.000*
$\hat{\sigma}^2(B\pi)$	96	0.024
$\hat{\sigma}^2(\gamma)$	2	0.070
$\hat{\sigma}^2(\alpha\gamma)$	14	0.000*
$\hat{\sigma}^2(\gamma\pi)$	192	0.073
$\hat{\sigma}^2(\theta)$	3	0.008
$\hat{\sigma}^2(\alpha\theta)$	21	0.000*
$\hat{\sigma}^2(\theta\pi)$	288	0.002
$\hat{\sigma}^2(B\gamma)$	2	0.010
$\hat{\sigma}^2(\alpha B\gamma)$	14	0.056
$\hat{\sigma}^2(B\gamma\pi)$	192	0.339
$\hat{\sigma}^2(B\theta)$	3	0.000*
$\hat{\sigma}^2(\alpha B\theta)$	21	0.002
$\hat{\sigma}^2(B\theta\pi)$	288	0.021
$\hat{\sigma}^2(\gamma\theta)$	6	0.003
$\hat{\sigma}^2(\alpha\gamma\theta)$	42	0.000*
$\hat{\sigma}^2(\gamma\theta\pi)$	576	0.007
$\hat{\sigma}^2(B\gamma\theta)$	6	0.006
$\hat{\sigma}^2(\alpha B\gamma\theta)$	42	0.017
$\hat{\sigma}^2(B\gamma\theta\pi)$	576	0.235

*Negative estimate has been replaced by zero

Note: α = classes, π = students, B = modes, γ = occasions, θ = raters.

Test of Homoscedasticity Assumption

The generalizability coefficients obtained from the intraclass correlation formulas are unbiased only if homogeneity of variance assumptions are met. To test this assumption in the context of the split-plot factorial design, a procedure described by Kirk (1968), pp.258-261) was used. The test for the equality of the eight variance-covariance matrices (corresponding to the eight classes) resulted in a chi-square value of 35.11. With 2100 degrees of freedom, the observed chi-square was not significant at the .10 level.

Since the eight matrices were not significantly different, a pooled variance-covariance matrix was constructed. Testing for the equality of the diagonal elements in the pooled matrix resulted in a chi-square value of 30.78, which was not statistically significant at the .10 level with 23 degrees of freedom. This result indicated that differences among the diagonal elements in the pooled matrix were not statistically significant. The results from these two tests lent support to the homogeneity of variance assumption.

Generalizability Coefficients

The coefficients reported in this section were obtained by substituting the point estimates from Table 4 into the formulas derived in Table 2. Forty nine coefficients were estimated, corresponding to seven different universes of generalization and seven different combinations of condition frequency. These coefficients are reported in Table 5. The first five represent combinations which yield a total of 24 observations on each person. Within that restriction, the combinations are included to show which facet needs to be sampled most frequently.

TABLE 5

GENERALIZABILITY COEFFICIENTS FOR SEVEN UNIVERSES OF GENERALIZATION AND SELECTED CONDITION COMBINATIONS

Universe of Generalization	CONDITION COMBINATIONS						
	4 Raters 2 Modes 3 Occasions	2 Raters 4 Modes 3 Occasions	4 Raters 1 Mode 6 Occasions	2 Raters 2 Modes 6 Occasions	4 Raters 6 Modes 1 Occasion	1 Rater 2 Modes 2 Occasions	1 Rater 1 Mode 1 Occasion
R Infinite M Infinite O Infinite	.765	.825	.761	.834	.703	.624	.330
R Fixed M Infinite O Infinite	.766	.829	.762	.836	.704	.628	.332
R Infinite M Fixed O Infinite	.791	.840	.814	.862	.711	.646	.353
R Infinite M Infinite O Fixed	.819	.883	.788	.863	.851	.690	.400
R Fixed M Fixed O Infinite	.798	.848	.879	.890	.714	.669	.375
R Fixed M Infinite O Fixed	.822	.889	.790	.867	.855	.700	.409
R Infinite M Fixed O Fixed	.970	.965	.965	.960	.973	.865	.747

The last two combinations are included to show the effect on the coefficients of minimum sampling.

The smallest coefficient obtained, .330, corresponds to a situation where generalization is intended across raters, modes, and occasions but each facet is sampled only once. This situation may occur if a classroom teacher were to base the student's writing scores for the year on one sample of writing.

For the same universe of generalization, increasing the number of conditions for the mode and occasion facets by one, results in an increased coefficient of .624. The highest coefficient for that universe, .834, is obtained when six conditions for the occasion facet are sampled and the rater and mode facets are each sampled twice.

As the universe of generalization is restricted, by fixing one or more facets, the generalizability coefficients tend to increase. In all universes, the smallest coefficients are found when only one condition of each facet is sampled.

In the three universes having only one facet fixed, the highest coefficients correspond to the two situations where the mode by occasion combinations are sampled the most. In the last universe, where generalization is intended across raters only, all the coefficients are high.

The Error Variance $\sigma^2(\Delta)$

The variance components were also used in estimating the error variance $\sigma^2(\Delta)$, the square root of which may be used for obtaining confidence intervals around an individual's universe score. Several components, $\sigma^2(\Delta)$, were estimated corresponding to the seven different universes of generalization. The results of this estimation are presented in Table 6. For each universe of generalization, seven

TABLE 6

ESTIMATES OF THE ERROR VARIANCE $\sigma^2(\Delta)$ FOR
SEVEN UNIVERSES OF GENERALIZATION AND SELECTED CONDITION COMBINATIONS

Universe of Generalization	CONDITION COMBINATIONS						
	4 Raters 2 Modes 3 Occasions	2 Raters 4 Modes 3 Occasions	4 Raters 1 Mode 6 Occasions	2 Raters 2 Modes 6 Occasions	4 Raters 6 Modes 1 Occasion	1 Rater 2 Modes 2 Occasions	1 Rater 1 Mode 1 Occasion
R Infinite M Infinite O Infinite	.134	.101	.124	.086	.222	.257	.798
R Fixed M Infinite O Infinite	.130	.096	.121	.081	.219	.247	.788
R Infinite M Fixed O Infinite	.121	.095	.100	.074	.218	.245	.774
R Infinite M Infinite O Fixed	.086	.054	.078	.062	.078	.186	.655
R Fixed M Fixed O Infinite	.116	.087	.092	.064	.214	.225	.743
R Fixed M Infinite O Fixed	.083	.048	.097	.056	.073	.170	.635
R Infinite M Fixed O Fixed	.016	.020	.018	.021	.016	.087	.282

estimates are included. These estimates correspond to different sampling combinations. The first five combinations yield a total of 24 observations. The last two represent minimal sampling of conditions within each facet.

As shown on the table for the first three universes, and again for the fifth, in the column where two raters, two modes, and six occasions are sampled, the error variance is at a minimum. For the fourth and sixth universes, the second combination is the one which minimizes $\hat{\sigma}^2(\Delta)$. In the last universe, the first five combinations yield small error variances.

Supplementary Analysis

Since five of the seven negative estimates obtained were associated with the classes effect, α , a follow-up analysis was done eliminating the classes from the model. Dropping the classes resulted in a four-way balanced factorial design without replications. This was one of the designs considered by Medley and Mitzel (1963). For this design, the point estimates of the variance components were obtained using the mean squares from the analysis of variance reported in Table 7. These mean squares were substituted into the formulas for the point estimates given by Medley and Mitzel (1963, p.312).

The resulting estimates of the variance components are reported in Table 8. As shown in Table 8, three of the 15 point estimates were negative and have been replaced by zeros. Of these, only the estimate of the student-by-rater interaction had been positive in Table 4. The ratio of negative estimates to the total number of estimates is smaller for the model without the classes effects than for the initial model

TABLE 7

ANALYSIS OF VARIANCE
FROM A FOUR-WAY FACTORIAL
DESIGN WITHOUT REPLICATIONS
STUDENT(S) X MODE(M) X OCCASION(O) X RATER(R)

Source	df	SS	MS
S	103	1142.285	11.090
M	1	0.673	0.673
O	2	150.337	75.169
R	3	12.956	4.319
S x M	103	210.202	2.041
S x O	206	500.079	2.428
S x R	309	100.335	.325
M x O	2	21.073	10.536
M x R	3	0.149	0.050
O x R	6	8.211	1.368
S x M x O	206	387.677	1.882
S x M x R	309	101.143	0.327
S x O x R	618	161.372	0.261
M x O x R	6	5.821	0.970
S x M x O x R	618	152.762	0.247

TABLE 8

POINT ESTIMATES OF THE VARIANCE COMPONENTS FOR
THE FOUR-WAY FACTORIAL WITHOUT REPLICATIONS

VARIANCE COMPONENT*	POINT ESTIMATE
$\hat{\sigma}^2(\pi)$	0.355
$\hat{\sigma}^2(B)$	0.000**
$\hat{\sigma}^2(B\pi)$	0.006
$\hat{\sigma}^2(\gamma)$	0.076
$\hat{\sigma}^2(\gamma\pi)$	0.066
$\hat{\sigma}^2(\theta)$	0.006
$\hat{\sigma}^2(\theta\pi)$	0.000**
$\hat{\sigma}^2(B\gamma)$	0.019
$\hat{\sigma}^2(B\gamma\pi)$	0.409
$\hat{\sigma}^2(B\theta)$	0.000**
$\hat{\sigma}^2(B\theta\pi)$	0.027
$\hat{\sigma}^2(\gamma\theta)$	0.002
$\hat{\sigma}^2(\gamma\theta\pi)$	0.007
$\hat{\sigma}^2(B\gamma\theta)$	0.007
$\hat{\sigma}^2(B\gamma\theta\pi)$	0.007

*The same notation used for model (16) will be used here.

**Negative estimate has been replaced by zero.

including the classes effects. Therefore, using negative estimates as the criterion, it appears that eliminating the effects involving classes, α , from model (16) results in a better model for these data.

The estimates in Table 4, obtained by the MIVQUEO method, and those in Table 8 obtained through the analysis of variance mean squares, are very close. The similarity between the estimates obtained from the two different methods lends support to the validity of the MIVQUE as a useful method when the data are unbalanced. The analyses of variance approach, as was mentioned earlier, is universally accepted as the best method for balanced data.

Summary

Point estimates for all variance components in the model were obtained and reported in Table 4. Negative estimates were replaced by zeros. The magnitude of the estimates indicated that students could be differentiated on the basis of their ratings. However, the classes as units could not be distinguished. Of the three sources of error examined, the occasion facet constituted the greatest source. The mode facet was next in magnitude. Raters represented an insignificant source of errors.

The tests of homogeneity of variance lent support to the assumption that the variances within each condition combination were equal. Assuming homogeneity of variance, unbiased generalizability coefficients were obtained for seven universes of generalization. These universes represented generalization across one facet, two facets, or all three facets simultaneously. For each universe, seven coefficients were computed for possible D-studies with various combinations of condition frequencies. For most universes, the coefficients indicated that to

obtain acceptable levels of generalizability at least six samples of writing from each person are necessary. The only exception was when generalization was intended across raters only. The results for the error corresponding to the standard error of measurement, were similar to those based on the generalizability coefficients.

A supplementary analysis which compared the estimates obtained through the MIVQUE method to those derived using expected mean squares, resulted in similar values for all estimates in a model without the classes effect. These results were interpreted as lending support to the validity of the MIVQUE method.

CHAPTER V

DISCUSSION

In this study, generalizability theory was applied to the assessment of writing ability in young children. A universe of generalization was defined in terms of three facets: modes, occasions, and raters. Samples of children's writing performance were obtained under selected conditions from each facet. The design permitted the investigation of three main sources of error and their interactions. These sources were considered to affect the inference of writing ability from writing performance. Formulas for generalizability coefficients were derived for seven universes of generalization.

The first two sources of error were defined in terms of variability in the quality of the writing samples. This variability may result from changes in the subject's performance across time (occasions) and across assignment (modes). The third source of error may result from differences in the standard of judgement used by different raters when scoring the samples. Using the principles of generalizability theory, the relative contributions of these sources of error were examined via estimates of the variance components. The discussion of the results is focused on the interpretation of the variance components and the usefulness of the theory. The limitations of this and similar studies are also considered.

Interpretation of Variance Components

The largest component of variance was that associated with the students, indicating that it was possible to rank order the students on the basis of their ratings. This component represented the universe score variance. The classes component, on the other hand, was considered to be zero (the actual estimate was negative), indicating that the eight classes could not be differentiated as units on the basis of the ratings received by the students. All but three components of interactions involving the classes were also zero. The three non-zero components were: the class-by-mode-by-occasion, .056; the class-by-mode-by-rater, .002; and the interaction of the classes with all three facets, .017.

Generalization Across One Facet

The point estimates for the student-by-facet interactions for the mode, occasion, and rater facets were .024, .073, and .002, respectively. These interaction components reflect the relative contribution of each source of error when generalization is done along that one dimension only. No interaction would mean that students are similarly rank ordered across all conditions of that facet, thus generalization across all conditions would be possible. On comparing these three estimates, it appears that occasions represented the greatest source of error while raters represented the smallest. The large relative contribution of occasions to error means that students are not ranked in the same manner for all three occasions. Differential learning might have taken place during the school year. An implication is that when making an assessment of writing ability, it is important to note when, during the year, the

measure was obtained. If generalization is intended across different occasion conditions, then several conditions should be sampled.

The small component associated with the student-by-rater interaction indicates that the four raters ranked the students similarly. It seems possible, then, to train raters in applying the general impression scoring method systematically. Since this scoring method is both fast and efficient, large scale projects could confidently take advantage of it. It is important to remember that, after scoring several papers, the raters discussed those samples which received differing scores. Thus, it is not surprising that this source of error was minimal.

The student-by-mode interaction was large enough to indicate that changes in the task may result in different rankings of students. Different modes of writing may demand different abilities from the students. A piece of creative writing, for example, would require an exercise of the imagination while writing a report would require the ability to organize facts in a meaningful fashion.

The three main effect components associated with modes, occasions, and raters were .000, .070, and .008, respectively. These components reflect systematic changes and contribute to error only if absolute decisions are being made or when different conditions are sampled for different students. Again, the occasion component is the largest, indicating that the overall ratings were greater on some occasions than in others. It is possible that all students improved their writing performance during the school year. The rater component is small but higher than the student-by-rater interaction. This component

reflects any systematic rater bias. There appeared to be no systematic variability due to modes.

Generalization Across Two or Three Facets

When generalization is intended across more than one dimension, in addition to the components discussed in the previous section, those components involving the interactions among facets must be considered. The three-way interaction components involving the students and two facets were .339, .021, and .007 for the mode-occasion, mode-rater, and occasion-rater combinations, respectively. The first one is relatively large, almost equal in magnitude to the student component. The interpretation of that component is that differences is students' ranking across the mode conditions change as a function of the occasion conditions. A large component indicates that, when generalization is intended across these two facets, the conditions should be sampled frequently, if error is to be minimized. This fact is reflected in Table 7 where coefficients of generalizability are shown for several condition combinations. The largest coefficients correspond to situations where modes and occasions are sampled most frequently.

The student-by-mode-by-rater component reflects some variability due to differential ranking of students by the raters as a function of the mode. That is, raters were not as consistent in one mode as they were in the other. The small student-by-occasion-by-rater interaction indicates that raters were almost as consistent in one occasion as they were in the others.

The two-way interaction components among facets were .010, .000, and .003 for the mode-by-occasion, mode-by-rater, and occasion-by-rater components, respectively. These components enter into the error variance $\sigma^2(\Delta)$ but not $\sigma^2(\delta)$. Of these, only the mode-by-occasion component is large enough to warrant consideration. This component indicates that differences in the overall ratings across modes vary as a function of the occasion. For example, it is possible that all students performed better when writing the creative story at the beginning of the year. On the other hand, at the end of the year they might have done a better job on the factual reports. If all students had more practice in one mode during the year, their improved ability in that mode would be reflected in this component.

When generalizing across all three facets, two additional components of variance must be considered. The four-way interaction component involving students and all three facets was relatively large, .235. Since there were no replications within any three facet combination, this component was confounded with the error of replication. The magnitude of this component indicates that generalization across all three facets requires that more than one condition of at least one facet be sampled in order to minimize the error. The three-way interaction component among the three facets was relatively small, .006.

Based on the previous discussion, it may be concluded that the occasion facet represented a greater source of error than the mode facet. The mode facet, in turn, represented a greater source of error than the rater facet. With proper training and practice, the rater facet may be almost irrelevant. These findings agree with those of

Finlayson (1951) and Vernon and Millican (1954) who concluded that differences in essays contributed more to unreliability than differences in raters. The differences in essay were further investigated in this study, since essays were characterized along two dimensions. Both of those dimensions were found to be important in this study. Furthermore, one of them was found to be more important than the other.

These findings also support the recommendations made by experts in the field of language arts and discussed in Chapter II. To obtain a reliable assessment of writing ability more than one sample of writing should be collected on more than one occasion and on more than one mode. How many is more than one? That depends on the intended universe of generalization.

An examination of Tables 7 and 8 provides some guidelines for answering that question. In those tables seven universes of generalization are considered. The first universe represents generalization across all three facets. The next three reflect generalization across two facets only, the third facet is held constant. The last three universes correspond to generalization across one facet only: occasions, modes, and raters, in that order. Several condition combinations are included in each table.

The entries in Table 7 represent generalizability coefficients obtained via intraclass correlation formulas. The error variance entering into those coefficients is $\sigma^2(\delta)$. In general, the highest coefficients, across all seven universes, correspond to situations where 12 writing samples are collected (the second and fourth condition combinations). Collecting six writing samples (first, third, and fifth condition combinations) results in a decrease in the coefficients.

However, the decrease is not too drastic, except perhaps in situations where all six samples are collected in one occasion. This situation seems unrealistic since, in this case, writer fatigue would interfere with writing ability. If only four samples are collected and only one rater is used, the coefficients drop below .7 for most universes. With only one sample, as shown in the last condition combination, most coefficients would be unacceptable.

The entries in Table 8 represent the estimates of the error variance $\sigma^2(\Delta)$ which takes into account systematic effects. The square root of the entries, $\sigma(\Delta)$, represents the standard error of measurement. Thus, the information in Table 8 may be used in constructing confidence intervals around individuals' true scores. In general, the conclusions that may be made based on the results shown in this table are similar to those based on Table 7. That is, for these estimates, those condition combinations which maximize $\rho^2(x,\mu)$, also minimize $\sigma^2(\Delta)$.

Usefulness of Generalizability Theory

On the basis of this study it may be said that generalizability theory provides a useful method for estimating the reliability of measures of writing ability. With a clear definition of error and using repeated studies, it might have been possible to examine certain reliabilities of essay using classical methods. Those reliabilities which include components of interactions among facets would, of course, be impossible to obtain under classical methods. For those reliabilities which are estimable under classical methods, the treatment would be more awkward. The basic requirement under the framework of generalizability theory is that the source of error be identified as a facet and

that conditions of that facet be sampled and incorporated into the design. In that manner, the components of variance associated with that source are estimable. Including facets in a design is a popular method of control in educational research since, typically, this kind of research takes place in the natural setting. It follows that generalizability theory provides a practical methodology in those situations.

Given the applicability of the theory to problems of reliability, it is surprising that applications of it are scarce in the literature. Some possible explanations of this situation are considered here. These are: (a) the unfamiliarity of applied educational researchers with the methods, (b) the unavailability of formulas for more complex designs, or (c) the limitation imposed by the restriction of balance.

This application of the theory is a step in making the methods more familiar to a wider group of applied educational researchers. In particular, researchers in the field of compositional writing have been provided with estimates of variance components which may be useful in the planning of both comparative and absolute D-studies in that area. In addition, formulas for the generalizability coefficients have been derived for the design used in this study. Those formulas may be adapted to fit other designs which represent subsets of our universe of admissible observations. All that would be required is that those terms involving facets not included in the design be dropped from the formula.

As was demonstrated in this study, the restriction of balance is not necessary. Several methods are available for the estimation of variance components in unbalanced designs. One of those methods was

used in this study. Computer programs in SAS may be used to obtain the point estimates. The procedure available in the 1976 version of SAS uses Henderson's method 3. A future version of SAS will include, in addition to the current method, the MIVQUEO method which was used in this study. The point estimates obtained from the MIVQUEO method were very similar to those obtained for a reduced model via expected mean squares. These results were presented in the supplementary analysis of the previous chapter. Future research should focus on comparing the "goodness" of these different methods when applied to specific situations.

These computer programs have certain limitations when large design matrices are involved. For large design matrices, such as the one used in this study, the current SAS program requires an excessive amount of computer space and time. For example, approximately five hours would have been required to get the point estimates for the components in this study under the current version. The MIVQUEO method uses less time and memory but for large design matrices it still represents an expensive process.

However, the estimates of the variance components from one G-study may be used in subsequent D-studies involving a similar population of individuals and similar facets. The estimates computed for this study may be useful to persons working with fourth grade students of similar characteristics. A limitation is introduced by the high rate of attenuation in this sample. To the extent that the final sample is representative of the fourth grade population, our estimates are useful.

An additional limitation of this study is introduced by the small number of conditions sampled within each facet. As has been pointed out by Henderson, among others, the sampling error of the estimates of variance components is large when few conditions are used in the estimate.

On a different application of this design, then, it is possible that the estimates obtained would vary from the ones in this study. As the number of degrees of freedom increases, the accuracy of the estimate also increases. It should be noted that the components used in the generalizability coefficients have large numbers of degrees of freedom since they involve the student effect.

Summary and Conclusions

This study examined the problem of reliability of measures of writing ability in the context of generalizability theory. Three main sources of error variance were considered: raters, modes, and occasions. It may be concluded that errors resulting from variability in the quality of writing across occasions and modes outweigh those stemming from differences among raters. With training and practice, raters can consistently score the writing samples of students using a general impression method. This method proved to be both fast and easy to use. To improve the reliability of measures of written composition and decrease the standard error of measurement, the emphasis should be placed on collecting several samples of writing. On the basis of the estimates obtained in this study, collecting less than six samples would result in coefficients below .70. Assessing the reliability of measures of writing ability in terms of rater agreement, is skimming the problem. It is unfortunate that this issue is most commonly addressed in terms of inter-rater reliability.

This study demonstrated the potential of generalizability theory for clarifying problems of reliability. In applying the theory, the careful identification of potential sources of error is required. Also, consideration must be given to the type of inference which is to be made

from the observations. On the basis of these considerations, the universe of observations is defined. A carefully designed study will allow the estimation of all sources of error variance identified. As was shown in this study, it is not necessary to limit applications of the theory to balanced designs. Methods of variance component estimation for unbalanced designs are documented in the statistical literature and available in SAS, a popular package of statistical programs.

REFERENCES

Anderson, H. E., & Bashaw, W. L. An experimental study of first grade theme writing. American Educational Research Journal, 1968, 5, 239-247.

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. A user's guide to SAS. Raleigh, N. C.: SAS Institute, 1976.

Bortz, D. E. The written language patterns of intermediate grade children when writing compositions in three forms: descriptive, expository, and narrative. Dissertation Abstracts International, 1970, 30, 5332-A.

Box, G. E. P. Problems in the analysis of growth and wear curves. Biometrics, 1950, 6, 362-389.

Braddock, R. Evaluation of writing tests. In A. H. Grommon (Ed.), Reviews of selected published tests in English. Urbana: National Council of Teachers of English, 1976.

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in written composition. Champaign, Ill.: National Council of Teachers of English, 1963.

Brennan, R. L. The calculation of reliability from a split-plot factorial design. Educational and Psychological Measurement, 1975, 35(4), 779-788.

Burt, C. Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 1955, 8(2), 103-118.

Coffman, W. E. Essay examinations. In R. L. Thorndike (Ed.), Educational Measurement. Washington D. C.: American Council on Education, 1971.

Coffman, W. E., & Kurfman, D. A comparison of two methods of reading essay examinations. American Educational Research Journal, 1968, 5, 99-107.

Cohen, A. M. Assessing college students' ability to write compositions. Research in the Teaching of English, 1973, 7, 356-371.

Cooper, C. R., & Odell, L. (Eds.). Evaluating Writing: describing, measuring, judging. Urbana, Ill.: National Council of Teachers of English, 1977.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley & Sons, 1972.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: a liberization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Diederich, P. The problem of grading essays. Princeton: Educational Testing Service, 1957.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Fagan, W. T., Cooper, C. R., & Jensen, J. M. Measures for research and evaluation in the English language arts. Urbana, Ill.: National Council of Teachers of English, 1975.

Finlayson, D. S. The reliability of marking essays. British Journal of Educational Psychology, 1951, 21, 126-134.

Fisher, R. A. Statistical Methods for Research Workers. London: Oliver & Boyd, 1925.

French, J. W. Schools of thought in judging excellence of English themes. Proceedings of Invitational Conference on Testing Problems, Princeton: Educational Testing Service, 1962.

Gillmore, G. M., Kane, M., & Naccarato, R. W. The generalizability of student ratings of instruction: estimation of the teacher and course components. Journal of Educational Measurement, 1978, 15(1), 1-14.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Goodnight, J. H. Personal communication, June 14, 1978.

Gulliksen, H. Theory of Mental Tests, New York: Wiley & Sons, 1950.

Guttman, L. A special review of Harold Gulliksen, Theory of Mental Tests. Psychometrika, 1953, 18, 123-130.

Henderson, C. R. Estimation of variance and covariance components. Biometrics, 1953, 9, 226-252.

Hoyt, C. J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Johnson, L. V. Children's writing in three forms of composition. Elementary English, 1967, 44, 265-269.

Kane, M. T., & Brennan, R. L. The generalizability of class means. Review of Educational Research, 1977, 47(2), 267-292.

Kirk, R. Experimental design: procedures for the behavioral sciences. Belmont, Ca.: Brooks/Cole, 1968.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Levy, P. Generalizability studies in clinical settings. British Journal of Social and Clinical Psychology, 1974, 13, 161-172.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton-Mifflin, 1953.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Tests Scores. Reading, Mass.: Addison-Wesley, 1968.

Lloyd-Jones, R. Primary Trait Scoring. In C. R. Cooper, & L. Odell (Eds.), Evaluating Writing: describing, measuring, judging. Urbana, Ill.: National Council of Teachers of English, 1977.

Magnusson, P. Test theory. Reading, Ma.: Addison-Wesley, 1967.

Maxwell, A. E., & Pilliner, A. E. G. Deriving coefficients of reliability and agreement for ratings. British Journal of Mathematical and Statistical Psychology, 1968, 21, 105-116.

McCaig, R. A. What your director of instruction needs to know about standardized English tests. Language Arts, 1977, 54, 491-495.

McColly, W. What does educational research say about the judging of writing ability? Journal of Educational Research, 1970, 64(4), 148-156.

Meckel, H. C. Research on teaching composition and literature. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Mellenbergh, G. J. The replicability of observational measures. Psychological Bulletin, 1977, 84, 378-384.

Mellon, J. C. National Assessment and the Teaching of English. Urbana, Ill.: National Council of Teachers of English, 1975.

Overall, J. E., & Spiegel, D. K. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

Perron, J. D. The impact of mode on written syntactic complexity. Athens: University of Georgia, 1976. (ERIC Document Reproduction Service No. ED 126 531)

Pilliner, A. E. G. The application of analysis of variance to problems of correlation. British Journal of Psychology, Statistical Section, 1952, 5(1), 31-38.

Pope, M. The syntax of fourth graders' narrative and explanatory speech. Research in the Teaching of English, 1974, 8, 219-227.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 1960, 25, 261-271.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. Generalizability of stratified-parallel tests. Psychometrika, 1965, 30, 39-56.

Rao, C. R. Estimation of variance and covariance components in linear models. Journal of the American Statistical Association, 1972, 67, 112-115.

Rao, C. R. Minimum variance quadratic unbiased estimation of variance components. Journal of Multivariate Analysis, 1971, 1, 445-456.

Rowley, G. L. The reliability of observational measures. American Educational Research Journal, 1976, 13(1), 51-59.

Scheffe, H. The analysis of variance. New York: Wiley & Sons, 1959.

Searle, S. R. Linear Models, New York: Wiley & Sons, 1971a.

Searle, S. R. Topics in variance component estimation. Biometrics, 1971b, 27, 1-76.

Seegars, J. C. Form of discourse and sentence structure. Elementary English Review, 1933, 10(3), 51-54.

Selvage, R. Comments on the ANOVA strategy for the computation of intraclass reliability. Educational and Psychological Measurement, 1976, 36(3), 605-609.

Singleton, D. J. The reliability of ratings of the essay portion of the Language Skills Examination. Dissertation Abstracts International, 1977, 37, 7710-A.

Stanley, J. C. Anova principles applied to the grading of essay tests. Journal of Experimental Education, 1962, 30, 279-283.

Suhor, C. Mass testing in composition: is it worth doing badly? New Orleans: New Orleans Public Schools, June 1977.

Vaughn, G. M., & Corballis, M. C. Beyond tests of significance: estimating strengths of effects in selected ANOVA designs. Psychological Bulletin, 1969, 72, 204-213.

Veal, L. R., & Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 1971, 5, 37-45.

Vernon, P. E., & Millican, G. D. A further study of the reliability of English essays. British Journal of Statistical Psychology, 1954, 7(2), 65-74.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

APPENDIX A

POINT ESTIMATES OF THE VARIANCE COMPONENTS AS LINEAR COMBINATIONS OF MEAN SQUARES FOR THE SPLIT-PLOT FACTORIAL DESIGN WITH BALANCED DATA

$$\hat{\sigma}^2(\alpha) = 1/n_{s(c)}n_m n_o n_r [\,MS(\alpha)-MS(\pi)-MS(\alpha B)-MS(\alpha\gamma)-MS(\alpha\theta)+MS(\pi B)$$
$$+MS(\pi\gamma)+MS(\pi\theta)+MS(\alpha B\gamma)+MS(\alpha B\theta)+MS(\alpha\gamma\theta)-MS(\pi B\gamma)-MS(\pi B\theta)$$
$$-MS(\pi\gamma\theta)-MS(\alpha B\gamma\theta)+MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\pi) = 1/n_m n_o n_r [\,MS(\pi)-MS(\pi B)-MS(\pi\gamma)-MS(\pi\theta)+MS(\pi B\gamma)+MS(\pi B\theta)$$
$$+MS(\pi\gamma\theta)-MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(B) = 1/n_{s(c)}n_c n_o n_r [MS(B)-MS(B\gamma)-MS(B\theta)+MS(B\gamma\theta)-MS(\alpha B)+MS(\alpha B\gamma)$$
$$+MS(\alpha B\theta)-MS(\alpha B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\alpha B) = 1/n_{s(c)}n_o n_r [MS(\alpha B)-MS(\alpha B\gamma)-MS(\alpha B\theta)+MS(\alpha B\gamma\theta)-MS(\pi B)+MS(\pi B\gamma)$$
$$+MS(\pi B\theta)-MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\pi B) = 1/n_o n_r [MS(\pi B)-MS(\pi B\gamma)-MS(\pi B\theta)+MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\gamma) = 1/n_{s(c)}n_c n_m n_r [MS(\gamma)-MS(\gamma B)-MS(\gamma\theta)+MS(B\gamma\theta)-MS(\alpha\gamma)+MS(\alpha\gamma\theta)$$
$$+MS(\alpha B\gamma)-BS(\alpha B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\alpha\gamma) = 1/n_{s(c)}n_m n_r [MS(\alpha\gamma)-MS(\alpha\gamma B)-MS(\alpha\gamma\theta)+MS(\alpha B\gamma\theta)-MS(\pi\gamma)+MS(\pi\gamma B)$$
$$+MS(\pi\gamma\theta)-MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\pi\gamma) = 1/n_m n_r [MS(\pi\gamma)-MS(\pi\gamma B)-MS(\pi\gamma\theta)+MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\theta) = 1/n_{s(c)}n_c n_m n_o [MS(\theta)-MS(\theta B)-MS(\theta\gamma)+MS(B\gamma\theta)-MS(\alpha\theta)+MS(\alpha\theta B)$$
$$+MS(\alpha\theta\gamma)-MS(\alpha B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\alpha\theta) = 1/n_{s(c)}n_m n_o [MS(\alpha\theta)-MS(\alpha\theta B)-MS(\alpha\theta\gamma)+MS(\alpha B\gamma\theta)-MS(\pi\theta)+MS(\pi\theta B)$$
$$+MS(\pi\theta\gamma)-MS(\pi B\gamma\theta)] \ .$$

$$\hat{\sigma}^2(\pi\theta) = 1/n_m n_o [MS(\pi\theta) - MS(\pi\theta B) - MS(\pi\theta\gamma) + MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(B\gamma) = 1/n_{s(c)} n_c n_r [MS(B\gamma) - MS(B\gamma\theta) - MS(\alpha B\gamma) + MS(\alpha B\gamma\theta)] .$$

$$\hat{\sigma}^2(\alpha B\gamma) = 1/n_{s(c)} n_r [MS(\alpha B\gamma) - MS(\alpha B\gamma\theta) - MS(\pi B\gamma) + MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(\pi B\gamma) = 1/n_r [MS(\pi B\gamma) - MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(B\theta) = 1/n_{s(c)} n_c n_o [MS(B\theta) - MS(B\gamma\theta) - MS(\alpha B\theta) + MS(\alpha B\gamma\theta)] .$$

$$\hat{\sigma}^2(\alpha B\theta) = 1/n_{s(c)} n_o [MS(\alpha B\theta) - MS(\alpha B\gamma\theta) - MS(\pi B\theta) + MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(\pi B\theta) = 1/n_o [MS(\pi B\theta) - MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(\gamma\theta) = 1/n_{s(c)} n_c n_m [MS(\gamma\theta) - MS(B\gamma\theta) - MS(\alpha\gamma\theta) + MS(\alpha B\gamma\theta)] .$$

$$\hat{\sigma}^2(\alpha\gamma\theta) = 1/n_{s(c)} n_m [MS(\alpha\gamma\theta) - MS(\alpha B\gamma\theta) - MS(\pi\gamma\theta) + MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(\pi\gamma\theta) = 1/n_m [MS(\pi\gamma\theta) - MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(B\gamma\theta) = 1/n_{s(c)} n_c [MS(B\gamma\theta) - MS(\alpha B\gamma\theta)] .$$

$$\hat{\sigma}^2(\alpha B\gamma\theta) = 1/n_{s(c)} [MS(\alpha B\gamma\theta) - MS(\pi B\gamma\theta)] .$$

$$\hat{\sigma}^2(\pi B\gamma\theta) = MS(\pi B\gamma\theta) .$$

BIOGRAPHICAL SKETCH

María Magdalena Llabre was born in Matanzas , Cuba, on September 22, 1950. She and her family immigrated to the United States in 1962.

Upon graduating from Miami Senior High School in 1969, she enrolled at the University of Florida where she received a Bachelor of Arts degree with a double major in Psychology and Mathematics. Following graduation María returned to Miami to teach mathematics at John F. Kennedy Junior High School for one year.

In 1974 she was admitted to the doctoral program in the Foundations of Education Department at the University of Florida. She received the M. A. E. degree in Educational Psychology in 1976.

While in graduate school, María worked in the evaluation of Project Follow Through and served as an evaluation consultant at P. K. Yonge Laboratory School. She was also a teaching assistant in research and statistics courses in the College of Education for three years.

She is currently a member of Phi Beta Kappa, the American Educational Research Association, the American Statistical Association, and the National Council for Measurement in Education.

María and her husband Brainard Hines will be moving to Miami where she has accepted a teaching position at the University of Miami starting in August, 1978.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

William B. Ware, Chairman
Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda M. Crocker
Associate Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Ramon C. Littell
Associate Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

John M. Newell
Professor of Foundations of Education

This dissertation was submitted to the Graduate Faculty of the Department of Foundations of Education in the College of Education and to the Graduate Council, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1978

Robert R. Sherman
Chairman, Foundations of Education

Dean, Graduate School